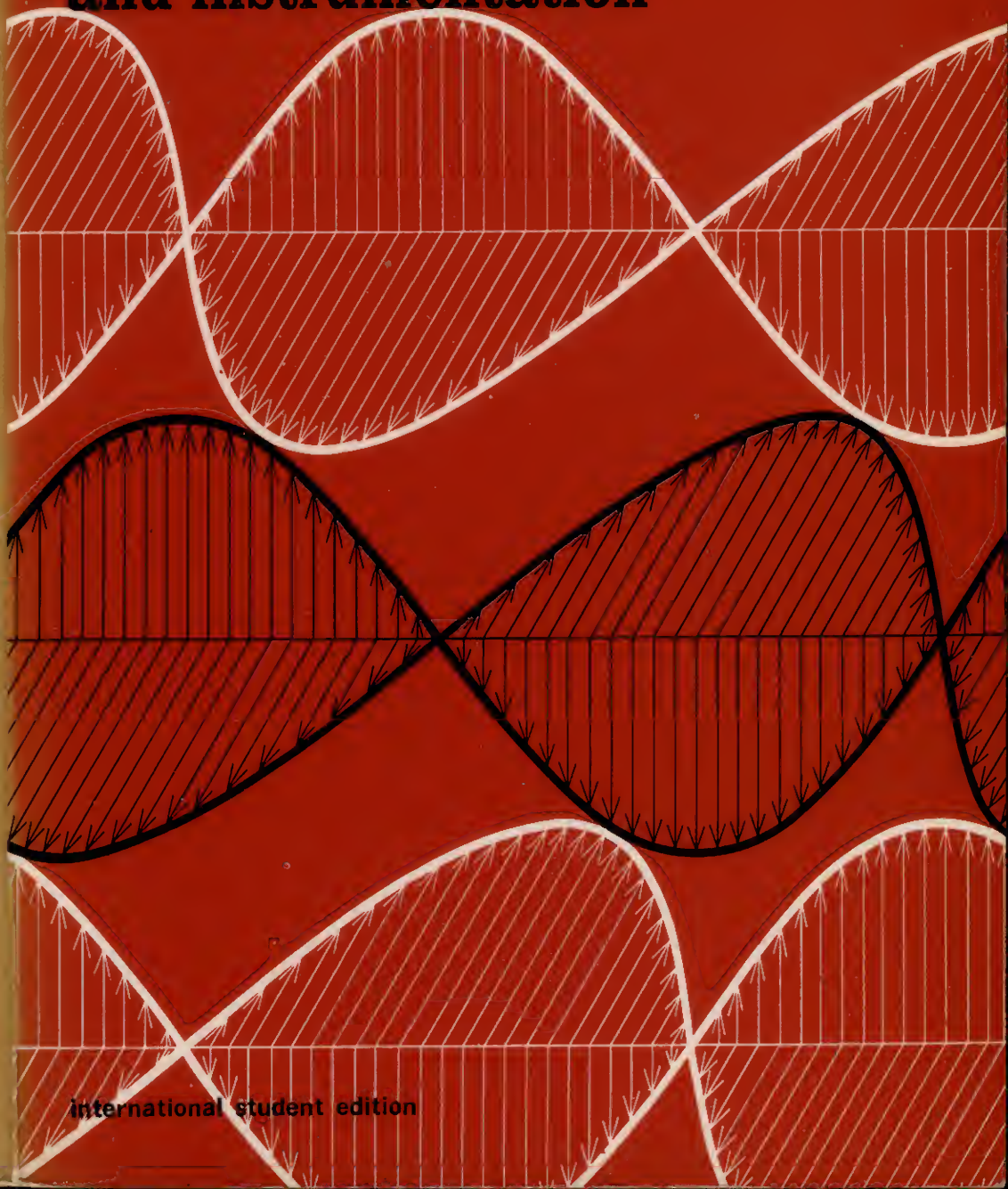
OLIVER ◆ CAGE

electronic measurements and instrumentation



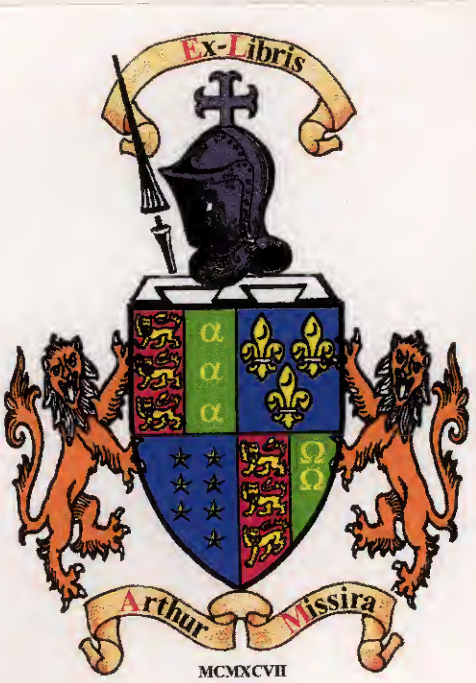
international student edition

electronic measurements and instrumentation

OLIVER ◆ CAGE

McGraw-Hill

R MW
TS6



ELECTRONIC MEASUREMENTS AND INSTRUMENTATION

INTER-UNIVERSITY ELECTRONICS SERIES

Consulting Editor

Charles Susskind, *University of California, Berkeley*

Advisory Committee

Frederick E. Terman, *Vice President, Emeritus, Stanford University*
Ernst Weber, *President, Polytechnic Institute of Brooklyn*
John R. Whinnery, *Professor of Electrical Engineering, University of California, Berkeley*

Steering Committee

Edward E. David, Jr., *Bell Telephone Laboratories*
Hubert Heffner, *Stanford University*
John G. Linvill, *Stanford University*
William R. Rambo, *Stanford University*
Mischa Schwartz, *Polytechnic Institute of Brooklyn*
John G. Truxal, *Polytechnic Institute of Brooklyn*
Lotfi A. Zadeh, *University of California, Berkeley*

Books in Series

Vol. 1: *Jamieson et al.* Infrared Physics and Engineering, 1964
Vol. 2: *Bennett and Davey.* Data Transmission, 1965
Vol. 3: *Sutton.* Direct Energy Conversion, 1966
Vol. 4: *Schwartz, Bennett, and Stein.* Communications Systems and Techniques, 1966
Vol. 5: *Luxenberg and Kuehn.* Display Systems Engineering, 1968
Vol. 6: *Balakrishnan.* Communication Theory, 1968
Vol. 7: *Collin and Zucker.* Antenna Theory, Part 1, 1969
Collin and Zucker. Antenna Theory, Part 2, 1969
Vol. 8: *Zadeh and Polak.* System Theory, 1969
Vol. 9: *Schwan.* Biological Engineering, 1969
Vol. 10: *Clynes and Milsum.* Biomedical Engineering Systems, 1969
Vol. 11: *Huelsman.* Active Filters: Lumped, Distributed, Integrated, Digital, and Parametric, 1970
Vol. 12: *Oliver and Cage.* Electronic Measurements and Instrumentation, 1971
Vol. 13: *Smit.* Magnetic Properties of Materials, 1971

INTER-UNIVERSITY ELECTRONICS SERIES, VOL. 12

ELECTRONIC MEASUREMENTS AND INSTRUMENTATION

EDITED BY

Bernard M. Oliver

Vice-President for Research and Development
Hewlett-Packard Company

John M. Cage

Hewlett-Packard Laboratories
Hewlett-Packard Company

INTERNATIONAL STUDENT EDITION

McGRAW-HILL INTERNATIONAL BOOK COMPANY

Auckland Bogotã Guatemala Hamburg Johannesburg Lisbon
London Madrid Mexico New Delhi Panama Paris San Juan
São Paulo Singapore Sydney Tokyo

ELECTRONIC MEASUREMENTS AND INSTRUMENTATION

INTERNATIONAL STUDENT EDITION

Exclusive rights by McGraw-Hill Kogakusha, Ltd., for manufacture and export. This book cannot be re-exported from the country to which it is consigned by McGraw-Hill.

V

Copyright © 1971 by McGraw-Hill, Inc. All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.
Library of Congress Catalog Card Number 71-124141

07-047650-0

When ordering this title use ISBN 0-07-085544-7

TOSHO PRINTING CO., LTD. TOKYO JAPAN

CONTENTS

Series Purpose xi
Contributors xiii
Preface xv
Further Preface xvii

One BASIC PRINCIPLES 1

1-1 The Role of Measurement 2
1-2 The Units of Measurement 6
1-3 Standards of Units of Measurement 11

Two SINE-WAVE TESTING OF LINEAR SYSTEMS 16

2-1 Mathematical Background 17
2-2 Gain or Loss Measurement 19
2-3 The Measurement of Phase 23
2-4 Automatic Network Analyzers 29
2-5 Measurement of Delay Distortion 30
2-6 The Measurement of Loop Gain 35
2-7 The Measurement of Nonlinearity 40
2-8 Precautions in Sine-wave Testing 42

Three SQUARE-WAVE AND PULSE TESTING OF LINEAR SYSTEMS 44

3-1 Tools and Techniques 46
3-2 Relations between Transient and Sinusoidal Responses 47
3-3 Response to Generalized Inputs 49
3-4 Effect of Low-end Cutoffs on Square-wave Response 57
3-5 Time-domain Reflectometry 61

Four MEASUREMENTS OF, WITH, AND IN THE PRESENCE OF NOISE 65

4-1 Mathematical Background 66
4-2 Measurement of Noise 80
4-3 Measurements with Noise as a Test Signal 87
4-4 Measurements by Pseudorandom Test Signals 93
4-5 Measurement in the Presence of Noise 97

Five SIGNAL ANALYSIS BY DIGITAL TECHNIQUES 106

5-1 Fourier Transform—Review of Basic Theory 108
5-2 Statistics—Review of Basic Theory 112
5-3 Signal Analysis 114
5-4 Summary 150

Six FREQUENCY AND TIME MEASUREMENTS 156

6-1 Time Definitions and Standards 157
6-2 Standard Frequency and Time-Signal Broadcasts 158
6-3 Time and Frequency Standards 162
6-4 Frequency Measuring Instruments 172
6-5 Frequency Synthesizers 179

Seven DIRECT-CURRENT INSTRUMENT AMPLIFIERS 186

7-1 Direct-current Amplifier Considerations 188
7-2 Direct-current Amplifier with Automatic Reset 197
7-3 Differential Amplifiers 201
7-4 Chopper Amplifiers 204

Eight VOLTAGE AND CURRENT MEASUREMENTS 210

8-1 Introduction to DVMs 211
8-2 Nonintegrating Types of DVMs 213
8-3 Digital Voltmeters with Counting Circuitry 219
8-4 Normal-mode Rejection 230
8-5 Common-mode Rejection 233
8-6 Principles of AC Voltage Measurements 236
8-7 Average-responding Detectors 239
8-8 Peak-responding Detectors 245
8-9 Peak-to-peak Detection 249
8-10 Root-mean-square–responding Detectors 250
8-11 Other Detection Methods 253
8-12 Sampling Voltmeters 255
8-13 Synchronous Detection 257
8-14 Direct-current Probes 259
8-15 Alternating-current Probe 261

Nine IMPEDANCE MEASUREMENT 264

9-1 Definitions and Formulas 264
9-2 Components and Standards 269
Resistors 269
Capacitors 272
Inductors 275
9-3 Meter Methods to Measure Impedance 276
Direct-current Meters 276
Capacitance and Inductance Meters 279
Complex Impedance Meters 280
Resistance and Impedance Comparators 281
9-4 Direct-current Bridges 282
The Wheatstone Bridge 282
Measurement of Low-valued Resistors 285
Measurement of High-valued Resistance 287
9-5 Low-frequency Bridges 288
General 288
Classification of Four-arm Bridges 291
Bridges with Inductively Coupled Ratio Arms 291
Special-purpose Bridges 297
Automatic and Semiautomatic Bridges 302
9-6 Radio-frequency Impedance Measurements 303
Problems at Radio Frequency 303
Radio-frequency Bridges 304
T Networks 305
Resonance Methods 308
The RF Meter Methods 310
9-7 Precision Measurements 311
Standardization of Impedance Units 311
Methods of Precision Measurement 313

Ten AUDIO FREQUENCY SIGNAL SOURCES 319

10-1 The Sinusoidal Audio-frequency Signal Source 319
10-2 The *LC* Oscillator 321
10-3 Resistance-Capacitance Signal Generators 328
10-4 The Wien Bridge Oscillator 331
10-5 The Practical Wien Bridge Oscillator 335
10-6 The Phase-shift Oscillator 337
10-7 The Ring Oscillator 338
10-8 The Beat-frequency Oscillator 340
10-9 The Polyphase Beat-frequency Oscillator 342
10-10 Sine-wave Synthesis 345

Eleven OSCILLOSCOPES 350

11-1 The Oscilloscope Function 351
11-2 Oscilloscope CRTs 353
11-3 Cathode-ray-tube Display-screen Characteristics 360

11-4 CRT Storage-target Characteristics 365
11-5 General-purpose Oscilloscopes 371
11-6 Sampling Oscilloscopes 398
11-7 Special-purpose Oscilloscopes 407
11-8 Accessories 418

Twelve RECORDERS 427

12-1 Galvanometric Recorders 428
12-2 Amplifiers for Galvanometric Recorders 440
12-3 Pen-driving Mechanisms 446
12-4 Servorecorders 456
12-5 Magnetic Recording 461
12-6 Magnetic Recording Techniques 464

Thirteen MEASUREMENTS ON AUDIO AND VIDEO AMPLIFIERS 480

13-1 Transfer Gain and Transfer Function 481
13-2 Steady-state Gain and Phase Measurement in Amplifiers 484
13-3 Gain Measurement with Extreme Accuracy 489
13-4 Common-mode Rejection and Input Balance 491
13-5 Common-mode Rejection and Balance with Transformer Coupling 495
13-6 Differential Amplifiers with Low CMR and Balance Ratios 496
13-7 Dynamic Range and Distortion 500
13-8 Slew Limiting 505

Fourteen MEASUREMENTS ON TRANSMITTERS AND RECEIVERS 507

14-1 General-performance Characteristics 508
14-2 Basic Measurements 510
14-3 Special System Measurements 510
14-4 Measurements on Receiving Systems 516
14-5 Sinad Sensitivity 519
14-6 Modulation-acceptance Bandwidth 520
14-7 Correlation of Sensitivity with Noise Figure 520
14-8 Automatic-gain-control Characteristics 523
14-9 Measurements on Transmitting Systems 525
14-10 Radio Equipment Specifications 538

Fifteen MICROWAVE SIGNAL SOURCES 544

15-1 Microwave Transistor Oscillators 545
15-2 Solid-state Microwave Amplifiers 550
15-3 Other Solid-state Microwave Sources 556

15-4 Solid-state Microwave Oscillators 565
15-5 Comparison of Solid-state Sources 572
15-6 Microwave-Signal Generators 576
15-7 Amplitude Modulators for Signal Generators 584
15-8 Microwave Sweep Generators 588

Sixteen MICROWAVE SIGNAL ANALYSIS 595

16-1 Power Measurement 595
16-2 Measurement of Power Exceeding 100 mW 604
16-3 Pulsed-power Measurements 605
16-4 Mismatch Considerations 607
16-5 Application Considerations 608
16-6 Microwave Frequency Meters or Wavemeters 611
16-7 Spectrum Analysis 616
16-8 The Swept Superheterodyne Spectrum Analyzer 620
16-9 Wave Analyzers 632
16-10 The TRF Spectrum Analyzer 633
16-11 Multifilter Real-time Spectrum Analyzer 634
16-12 The Tracking Generator Counter 635
16-13 Techniques and Applications for Analyzers 637
16-14 Some Rules of Thumb for Choosing Bandwidth and Other Control Settings when Viewing Pulsed RF Spectrums 651
16-15 Electromagnetic Interference Measurements 653

Seventeen MICROWAVE NETWORK ANALYSIS 654

17-1 Reflection and Impedance Measurement 655
17-2 Attenuation Measurements 659
17-3 Ferrite Devices 661
17-4 Two-port Network Theory 671
17-5 Theory of *s* Parameters 673
17-6 Network Calculations by Using Scattering Parameters 678
17-7 Measurement of *s* Parameters 686
17-8 Measurement Techniques 688

Eighteen AUTOMATED MEASUREMENT SYSTEMS 704

18-1 Low-level Multichannel Data Acquisition Systems 705
18-2 Common-mode Noise Rejection 707
18-3 Normal-mode Noise Rejection 709
18-4 Low-noise Preamplifiers 710
18-5 Crosstalk 710
18-6 Thermal Voltages 712
18-7 Scanners 713
18-8 Automatic Analyzer Systems 714
18-9 Automatic Test Systems 716

Index 721

SERIES PURPOSE

The explosive rate at which knowledge in electronics has expanded in recent years has produced the need for unified state-of-the-art presentations that give authoritative pictures of individual fields of electronics.

The Inter-University Electronics Series is designed to meet this need by providing volumes that deal with particular areas of electronics where up-to-date reference material is either inadequate or is not conveniently organized. Each volume covers an individual area, or a series of related areas. Emphasis is upon providing timely and comprehensive coverage that stresses general principles, and integrates the newer developments into the over-all picture. Each volume is edited by an authority in the field and is written by several coauthors, who are active participants in research or in educational programs dealing with the subject matter involved.

The volumes are written with a viewpoint and at a level that makes them suitable for reference use by research and development engineers and scientists in industry and by workers in governmental and university laboratories. They are also suitable for use as textbooks in specialized courses at graduate levels. The complete series of volumes will provide a reference library that should serve a wide spectrum of electronic engineers and scientists.

The organization and planning of the series is being carried out with the aid of a Steering Committee, which operates with the counsel of an

Advisory Committee. The Steering Committee concerns itself with the scope of the individual volumes and aids in the selection of editors for the different volumes. Each editor is in turn responsible for selecting his coauthors and deciding upon the detailed scope and content of his particular volume. Over-all management of the Series is in the hands of the Consulting Editor.

Frederick Emmons Terman

CONTRIBUTORS

STEPHEN F. ADAM *Hewlett-Packard Company, Palo Alto, California*

PAUL BAIRD *Hewlett-Packard Company, Loveland, Colorado*

ALAN S. BAGLEY *Manager, Santa Clara Division, Hewlett-Packard Company, Santa Clara, California*

ARNDT B. BERGH *Hewlett-Packard Company, Cupertino, California*

L. BESSER *Fairchild Electronics Corporation, Mountain View, California*

W. B. BRUENE *Collins Radio Company, Cedar Rapids, Iowa*

RODERICK CARLSON *Hewlett-Packard Company, Palo Alto, California*

ROBERT L. DUDLEY *Hewlett-Packard Company, Loveland, Colorado*

JOHN J. DUPRE *Hewlett-Packard Company, Palo Alto, California*

M. D. EWY *Hewlett-Packard Company, Palo Alto, California*

CHARLES O. FORGE *Durrum Instruments, Palo Alto, California*

DOUGLAS GRAY *Hewlett-Packard Company, Palo Alto, California*

HENRY P. HALL *General Radio Company, Concord, Massachusetts*

D. B. HALLOCK *Collins Radio Company, Cedar Rapids, Iowa*

STEVE HAMILTON *Hewlett-Packard Company, Palo Alto, California*

FRED HANSON *Hewlett-Packard Company, Loveland, Colorado*

CHARLES H. HOUSE *Hewlett-Packard Company, Colorado Springs, Colorado*

WILLIAM HEINZ *Hewlett-Packard Company, Palo Alto, California*

BILL KAY *Hewlett-Packard Company, Loveland, Colorado*

CHARLES KINGSFORD-SMITH *Hewlett-Packard Company, Loveland, Colorado*

WILLIAM McCULLOUGH *Hewlett-Packard Company, Loveland, Colorado*

C. D. MEE *International Business Machines Corporation*

ARTHUR MILLER *Consulting Engineer*

RICHARD Y. MOSS, II *Hewlett-Packard Company, Loveland, Colorado*

DONALD E. NORGAARD *Hewlett-Packard Company, Palo Alto, California*

BERNARD M. OLIVER *Hewlett-Packard Company, Palo Alto, California*

FRED PRAMANN *Hewlett-Packard Company, Palo Alto, California*

RONALD W. POTTER *Hewlett-Packard Company, Santa Clara, California*

WALLACE RASMUSSEN *Hewlett-Packard Company, Palo Alto, California*

GORDON ROBERTS *Engineering Manager, Hewlett-Packard, Ltd., South Queensferry, West Lothian, Scotland*

DOUGLAS K. RYTTING *Hewlett-Packard Company, Palo Alto, California*

OTTO S. TALLE, JR. *Hewlett-Packard Company, San Diego, California*

TERRY E. TUTTLE *Hewlett-Packard Company, Loveland, Colorado*

CRAIG WALTER *Hewlett-Packard Company, Loveland, Colorado*

LARRY A. WHATLEY *Hewlett-Packard Company, Loveland, Colorado*

PREFACE

This book is addressed to anyone with some knowledge of electricity, electronics, and circuit theory who wishes to become familiar with the great variety of electronic instruments and measuring systems available today and with the kinds of measurements they can make. Because the field has grown so big and exhibits such diversity, we have had to omit a great deal of material both of a fundamental nature and of a specialized nature. For example, there is no chapter on dc measurements as such, although many instruments used in dc measurement are described fully in other chapters. At the other extreme, tremendously complex data telemetry systems such as are used in the space program are not treated, though many of the components of such systems are. Thus, we have tried to steer a course that avoids both the obvious and the esoteric, hoping thereby to provide a more generally useful book with wide appeal.

Several years ago I was beguiled by the publishers into contracting to write this work. Their arguments were persuasive: the advent of the transistor had obsoleted existing books on the subject, the art had advanced greatly since these works were written, I had personally been involved with much of that advance, etc. Overcome by this suasion, I agreed to accept the assignment. My attitude at the time seemed euphoric, but now appears to me to have been more one of conceit, for as I approached my task closer, it loomed ever larger, and soon humbled me into a state of paralysis. After a long time during which I tried to delude myself that I could simultaneously write a book, be president of the IEEE, and carry out my regular job, a time during which my sense of guilt steadily grew, I had the good fortune to have John Cage offer his

services in recruiting other writers and in compiling and consolidating their efforts. This book never would have materialized without their contributions and without John's steadfast attention to details and to schedules.

It is not the book I would have written, but it is, I think, a good book and in many ways superior to what I might have produced, given the time. For the sake of all those who have contributed I hope you and many, many others will find it valuable.

B. M. Oliver

FURTHER PREFACE

As our technological civilization develops, both the amount and the complexity of measurement proliferates. I think it is axiomatic that the instrumentation art must grow *faster* than the science and engineering activities in which measurement is necessary. The mercurial growth of electronic measurement and instrumentation for use in electronics and in electrical engineering certainly supports the axiom.

This inevitable growth in instrumentation presents a problem to the engineering or science student, to the young practicing engineer whose skill in measuring his working parameters is already obsolescent, to the manager who is keenly aware of the role played by clever measurement techniques in technical progress. Before these people can participate in the surge of progress, I think they must learn about the principles and the creative combinations of ideas in modern instruments. Then they can use the instruments skillfully and even conceive *new* instruments. This book is intended to help these people.

Consider Chap. 4, "Measurements of, with, and in the Presence of Noise." Without an understanding of this material, how can an engineer or scientist penetrate very far in *any* discipline involving variable quantities? Or look at Chap. 5, "Signal Analysis by Digital Techniques." What an important instrument is one that adequately displays the Fourier transform of a function of time, even when the interval of integration is not infinity! Or Chap. 16: Do you really understand spectrum analyzers?

Several years ago, McGraw-Hill suggested that B. M. Oliver, Vice

President for Research and Development of the Hewlett-Packard Company, was one of the people who could be logically chosen to write a comprehensive reference book on electronic measurement and test instruments. However, it was apparent that the book would cover so many subdisciplines, require information from so many specialists, and require so many hours that it could not be fitted into Oliver's activities. It was suggested that I work with him to achieve an acceptable time schedule. I knew from experience that working with Barney Oliver in any capacity would keep the cerebral vascular system healthy and transform senescence, as Shakespeare said, into "the silver livery of advised age."

It has been equally rewarding to work with the many specialists whose names appear with the chapter headings. To combine the notes and papers of more than thirty-five experts with some degree of continuity was educational, to say the least. These people were busy (and often nonaspiring as authors) but real authorities all the same.

We lacked space to treat measurement specifically in the fields of medicine, chemistry, and other nonelectrical disciplines. These subjects will require separate books.

I shall not discuss the plan of the book further in the Preface. Chapters 1, 2, and 3 serve that purpose well. However, I must express my great appreciation to Miss Helen Azadkhanian for her very patient preparation of the manuscript, to Mrs. Downs for her help with some of the drafting problems, and to my wife, Mildred, who graces everything with beauty and love.

John M. Cage

ELECTRONIC MEASUREMENTS AND INSTRUMENTATION

CHAPTER ONE

BASIC PRINCIPLES

Bernard M. Oliver

Hewlett-Packard Company
Palo Alto, California

Although the measurement of simple physical quantities dates from ancient times, measurement as a precise art is only a few hundred years old, and many of the quantities we measure today were not even known to exist or were at best ill understood a century ago. Even so fundamental a dimension as time was measured extremely crudely with sand and water clocks until Galileo's observations on the pendulum suggested replacing these dissipative mechanisms with resonant systems in which cycles are counted. Since that time, clocks have not changed in principle, though their accuracy has been improved enormously as better and better resonant systems were discovered. Today we use the atomic resonances of cesium and hydrogen to measure time with an accuracy that corresponds to less than a one-second error in thirty thousand years. No other physical quantity can yet be measured with this precision. But while horology may hold the current accuracy record, other areas too have greatly benefited from the application of electronics to their measurement problems.

Electronic measurements are of two kinds: those made *of electronic*

quantities such as voltage, capacitance, or field strength, and those made *by electronic means* of other quantities such as pressure, temperature, or flow rate. Electronic instrumentation came of age in solving the measurement needs of electronics itself, but in its maturity it is proving remarkably adaptable to other fields. In this book we shall mainly consider electronic instrumentation as a tool of its own trade. This avoids a great deal of repetition and superficiality and at the same time represents very little loss of generality, since the first step in measuring a nonelectrical physical variable electronically is to convert the variable to an electrical quantity.

1-1 The Role of Measurement

Science and technology are so intertwined with measurement as to be totally inseparable from it. It is true that modern measuring instruments are one of the fruits of science, but it is equally true that *without the ability to measure, there would be no science.* When Lord Kelvin warned that knowledge not expressible in numbers was "of a meager and unsatisfactory kind," he was not expressing a fetish; he was identifying an essential aspect of scientific knowledge. The laws of physics are quantitative laws, and their validity can only be established by precise measurement. It is the insistence on quantitative agreement of theory with experimental fact that distinguishes science from philosophy.

The careful astronomical observations of Tycho Brahe and the brilliant analysis of his data by Johannes Kepler illustrate very dramatically the contribution of accurate measurement to scientific progress. Plato, and the Greek philosophers who followed him, believed that the heavenly bodies, being perfect, were composed of the quintessence (literally the fifth essence of matter as distinct from earth, fire, air, and water) and that their motions must be eternal and perfect. Certainly the stars moved in circles, and it was believed that the motions of the planets could be described by an appropriate combination of uniform circular motions. For two thousand years the resolution of planetary motions into circular components was considered the most important problem in astronomy. The heliocentric theory of Aristarchus of Samos (250 B.C.), the geocentric theory of Ptolemy (A.D. 150), and even the heliocentric theory of Copernicus (A.D. 1543), all adhered to the concept of circular motions. But even though the Copernican theory greatly simplified the Ptolemaic theory by eliminating the large epicycles that were really the result of the earth's own motion, neither theory predicted the exact positions of the planets at all times. The error in both theories was often as much as two degrees.

To Tycho Brahe, who was born shortly after Copernicus' death, two

degrees of error was intolerable. He decided that, before any correct theory could be discovered, the actual positions of the planets over many years would have to be measured with far greater accuracy than ever before. With the financial support of Frederick II of Denmark he built very large and rigid quadrants and other instruments for measuring angles. These he mounted on stable foundations in his observatory, which he named Uraniborg, or "castle of the heavens." Then he *calibrated* his instruments so that he could subtract their errors from his observations. For twenty years he recorded the positions of the planets. After the death of Frederick II he moved to Prague, where Kepler became his assistant.

Kepler was assigned the task of computing the orbit of Mars from Brahe's observations. After four years of arduous work Kepler came to a painful conclusion. No combination of the deferents and epicycles of the Copernican or the Ptolemaic systems would fit the facts. The motion of Mars could not be compounded out of regular circular motions as Plato had believed. The best solution Kepler found disagreed with observations by only eight minutes of arc. But Kepler knew that Tycho Brahe's observations could not be in error by more than two minutes of arc. With an integrity rare even in scientists, Kepler saw that beliefs twenty centuries old were doomed by an error only six minutes of arc too big to be allowable.

Kepler then went on to discover his famous laws of planetary motion. Eighty years later Newton showed that all these laws were a consequence of his own laws of motion and his theory of universal gravitation, and thus provided convincing proof of the latter. Shattered forever were the crystalline spheres that carried the planets in their Ptolemaic orbits. All the complex motions of the planets, which had puzzled men for ages, were distilled into one simple little equation.

Nor does the story end here, for later and much more accurate observations, with telescopes, showed that the orbit of Mercury precessed by 43 seconds of arc per century more than could be accounted for by perturbations of the other planets. This in turn later provided the best confirmation we yet have of Einstein's general theory of relativity, which subsumes Newton's law of gravitation as a special case.

The role of measurement in unraveling the mysteries of celestial mechanics is paralleled in other branches of science. Quantitative measurements of the stoichiometry of chemical reactions established the existence of the atom, and precise measurements in spectroscopy have helped reveal its structure. Today, measurements of the trajectories of nuclear fragments are gradually revealing the nature of the nucleus. X-ray diffraction studies have taught us how crystals are built and have provided important clues to the nature of deoxyribonucleic acid (DNA) and other

organic molecules. The list is endless, for after Tycho Brahe, Galileo, and Newton, science became experimental, and all experiments involve measurement. Man finally learned not to impose his beliefs on nature but, instead, humbly to ask questions of her and apply reason to her answers.

New discoveries in science provided new instruments for the study of nature and these studies produced new discoveries in a regenerative buildup that has been accelerating for the last two centuries and continues to accelerate today. Though much of physics has now been explored, many mysteries still remain at both extremes of size: the nucleus and the cosmos. The fields of particle physics and cosmology together with molecular biology are the major frontiers of modern science. All depend heavily upon instrumentation and measurement.

The science of optics produced the first major contributions to scientific instrumentation: the telescope, the microscope, and the spectroscope. When Galileo refined a Fleming's spyglass and turned it toward the heavens, a new era in astronomy was born. Later the spectroscope not only revealed new elements on earth, but provided the final, uncontrovertible proof that the stars themselves, like our sun, are composed of these same elements. The microscope showed the cellular structure of living matter and the microorganisms that are the cause of disease.

Imagine how different human history might have been had Aristarchus of Samos had a telescope and spectroscope, and Hippocrates a microscope! What Greek could have believed in the quintessence of matter having seen the mountains of the moon and spectral lines of earthly elements in sunlight? Or who could have insisted that all heavenly bodies revolved around the earth, having beheld the satellites of Jupiter? How could the deity have been so wasteful as to adorn the sky with stars not even visible to man's naked eye? What need for evil spirits if microbes cause disease? The impact of such discoveries, had they been made by the Greeks, would surely have greatly accelerated civilization and profoundly affected theology. Indeed the western world might have been spared the dark ages and the tortures of the Inquisition if only the Greeks had had better instrumentation.

In recent years both astronomy and biology have taken new leaps forward, again because of new tools, this time the result of progress in electronics. The radio telescope has enabled astronomers to study the matter between the stars in what was once thought of as simply space. Quasars, perhaps the most distant objects in the universe, and pulsars, believed to be star corpses composed almost entirely of neutrons, have been discovered with radio telescopes. Meanwhile, the electron microscope has revealed single strands of DNA and many of the fantastic transfer

mechanisms in the living cell that use the genetic code to construct proteins, antibodies, and enzymes. Living things too, it now seems certain, obey the laws of physics and chemistry.

The role of science is to discover the laws of nature and how they operate in complex systems. The role of engineering is to apply the discoveries of science to human needs. Scientists make discoveries that increase our understanding of the world. Engineers make inventions intended to increase our productivity (and thereby our standard of living), our mobility, and (it is hoped) our ability to survive. Instrumentation is a branch of engineering that serves not only science but all branches of engineering and medicine as well.

The precise measurement of dimensions, temperature, pressures, power, voltage, current, impedance, various properties of materials, and a host of other physical variables is as important to engineering as to science. Thus, mass production of goods that has produced our present affluent society would be impossible unless their parts could be made so nearly alike as to be completely interchangeable.

Eli Whitney, the inventor of the cotton gin, seems to be the first to have eliminated the need for selective assembly. In 1798 he obtained a contract to produce ten thousand muskets for the United States government and decided to "substitute correct and effective operations of machinery for that skill of an artist which is acquired only by long practice and experience." It took Whitney two years, during which time not a single gun was produced, to develop the machines, tools, and fixtures to do the job. Washington officials became nervous at the delay, but finally Whitney appeared before the Secretary of War and other Army officers with boxes containing all the parts of his musket. While they watched in amazement, Whitney assembled ten muskets, taking parts indiscriminately from the boxes. Afterward, in a letter to Monroe, Jefferson wrote: "He (Whitney) has invented molds and machines for making all the pieces of his locks as exactly equal, that take a hundred locks to pieces and mingle their parts and the hundred locks may be put together by taking the pieces that come to hand."

Accurate measurement is needed too for economy of design. A bridge several times stronger than needed to carry its heaviest possible load serves no one better and costs more than one designed to survive this worst load safely. For millions watching on television, the most dramatic moment of the Apollo 11 mission occurred when Neil Armstrong first set foot on the moon. But for many of the engineers who designed the vehicles and the computer programs, the most dramatic moment occurred two hours earlier when the lunar landing module set *its* feet on the moon. At that moment, only ten seconds worth of fuel remained. Close timing

indeed, and a tribute to the designers of the mission, for every pound of spare fuel that did not have to be allowed for in the landing module could be used to increase the payload of the lunar escape module.

Not only are instrumentation and measurement playing an increasingly important role in our technological society; electronics is playing an increasingly important role in instrumentation. The reasons for the latter are that most physical quantities can be converted by transducers into electrical signals and, once in this common form, they may be amplified, filtered, multiplexed, sampled, and measured. The measurements are easily obtained in or converted into digital form for automatic analysis and recording, or the data can be fed to servo systems for automatic process control. Electronic circuits are unexcelled in their ability to detect and amplify weak signals and in their ability to measure events of short duration. The incorporation of electronic sensors and circuits into instruments has vastly increased our ability to measure and thereby our ability to find nature's answers to new questions.

Where science will take us in the future, no one knows. That is what makes it such an exciting adventure. But one thing seems certain. If social or political or ecological catastrophe can be avoided, science will continue to probe with new and even more sensitive instruments while the riddles of matter, of the origin of the universe, and of life are being answered. Perhaps in time we may be able to construct a philosophy in total accord with all knowledge. Or perhaps, as is more likely, we shall no longer feel the need for philosophy. For what is philosophy but intellectual speculation turned into belief, and what place is there for speculation except to develop premises to be tested?

1-2 The Units of Measurement

Set into the stone wall of Saint Stephen's Cathedral in Vienna are two iron bars with protruding ends. One is about a yard long and the other is about a meter long, but they are much older than either of these units of measure. In medieval times Vienna was the western terminus of caravans that carried the trade from the East, and these bars were used to measure the width of silk cloth and other fabrics imported by the traders. The church in those days was the keeper of physical as well as spiritual standards and required the infidel to measure up to the former if not the latter.

The measurement of quantities important in trade, such as length, mass, and volume, is as old as civilization itself, but very few units of ancient measure have been preserved. Today no one knows the exact length of the stadium or of the cubit. So when we read that Eratosthenes, in the third century B.C., having measured the angle of the sun's rays in

Alexandria at the moment the sun was directly overhead in Syene, and knowing the north-and-south distance between these cities, computed the circumference of the earth as 250,000 stadia, we can only admire his genius, but we cannot check his result with certainty.

The measurement of angles is unique in that the unit is dimensionless; no standard is needed, but only a numerical convention. Perhaps this accounts for the longevity of the Babylonian system of angle measurement, which is still in use today. In keeping with their sexaguesimal (base 60) number system, the Babylonians divided the angle of an equilateral triangle into 60 parts to get the degree. The degree was then divided into 60 tiny, or *minute*, divisions, and these in turn were divided into 60 *second*-order minute divisions, today called simply *minutes* and *seconds*. It is a pity that the Babylonians did not divide the circle into 24 parts, as we now divide the day, to obtain their basic units for then these two traditional measures of time and angle, incompatible as their subdivisions are with the decimal system, would at least be consistent with each other and astronomers would not have to reckon with two kinds of minutes and two kinds of seconds.[1]

As experimental science developed in the eighteenth and nineteenth centuries, the need for commonly accepted units of measure began to be felt. Without such standardization the intercomparison of results by workers in different countries was much more difficult. After the metric system was adopted in France in 1799, its measures of length and mass were gradually accepted, along with the already established unit of time, the second, as the units in which scientific findings in mechanics were reported. Even though many laboratories used different metric units such as the meter-gram-second (mgs) system, or the millimeter-milligram-second system, the integral powers of 10 relating these units made conversion relatively easy. Gradually however, the centimeter-gram-second (cgs) system became the universally accepted standard in science in the late nineteenth and early twentieth centuries. Not only were units standardized, but so were the symbols and names for the units of the various physical quantities. This too helped make the equations of science a sort of universal language easily understood by scientists everywhere.

Electrical Units. The early history of electrical units was complicated by the fact that the relations between electrostatics and electromagnetics

[1] A new unit, the *neugrad* (or grad), equal to one four-hundredth of a circle, has been introduced in Europe. It is subdivided decimally, but this is also possible with degrees, and the virtue of dividing a right angle into 100 parts rather than 90 is anything but obvious. Indeed it is more awkward to have common angles such as 30° expressed as 33⅓ grad. If any change is to be made, let us choose the new unit to be 15°!

were not yet clearly understood. In both fields the importance of tying electrical units to the earlier, well-established mechanical units of work and force was appreciated. Thus workers in electrostatics chose the coulomb relation for the force between two point charges q_1 and q_2, i.e.:

$$F = \frac{q_1 q_2}{\epsilon r^2} \tag{1-2-1}$$

with ϵ taken to be equal to unity in vacuum as the starting point for defining the unit of charge. Two equal charges that produced 1 dyne of force at a distance of 1 cm were each a unit charge. Since ϵ was taken as unity in vacuum, it was generally regarded as dimensionless and this gave charge the dimensions of $\sqrt{ML^3/T^2} = M^{1/2}L^{3/2}T^{-1}$ in the electrostatic unit (esu) system. Then workers in electromagnetics followed a similar route. The unit magnetic-pole strength m_1 and m_2 was that which produced unit force at unit distance in the relation

$$F = \frac{m_1 m_2}{\mu r^2} \tag{1-2-2}$$

with $m_1 = m_2$ and $\mu = 1$ in vacuum. But since magnetic poles, unlike charge, cannot be isolated in nature, this equation was replaced by the theoretically equivalent relation involving currents I_1 and I_2:

$$\frac{F}{l} = \frac{2\mu I_1 I_2}{d} \tag{1-2-3}$$

for the force F per unit length l between two infinitely long parallel conductors separated by a distance d. Again because μ was considered dimensionless, current, defined by Eq. (1-2-3), was assigned the dimensions $\sqrt{ML/T^2} = M^{1/2}L^{1/2}T^{-1}$. Charge, being the product of current and time, therefore had the dimensions $M^{1/2}L^{1/2}$ in the electromagnetic unit (emu) system.

The dimensions of the esu unit of charge were thus L/T times the dimensions of the emu unit of charge, and the same was true for other units in the two systems. Because the number of units in a given quantity is inversely proportional to the size of the unit, we see that $q_{\text{emu}} = vq_{\text{esu}}$, where v is some velocity. Upon substitution of the actual magnitudes, v turned out to be the velocity of light, a fact which strongly suggested to Maxwell and others that light is an electromagnetic phenomenon.

While both the esu and emu systems were being used by theoreticians and scientific experimenters, still a third system of so-called practical units, comprising the volt, ampere, coulomb, and watt, was developed for use in engineering. In 1863 the British Association for the Advancement of Science, which played a leading role in the early standardization of all

basic units, defined certain of these practical units as decimal multiples of the emu units, and so they remain today.

The dimensional disagreement between the esu and emu units for the same quantities made it clear that ϵ in Eq. (1-2-1) and μ in Eqs. (1-2-2) and (1-2-3) must not be considered dimensionless, and if not dimensionless, why give them the value of unity for vacuum? This practice, while it may have accelerated Maxwell's unification of electric and magnetic theory, also led to much confusion by concealing the fundamental physical difference between the electric field strength E and the electric displacement density D and even more between the magnetic field strength H and the magnetic induction B. As soon as it was realized that there was no more reason for choosing ϵ and μ equal to unity than for choosing γ equal to unity in Newton's law of gravitation

$$F = \gamma \frac{M_1 M_2}{r^2} \tag{1-2-4}$$

(and thereby defining a unit of mass), the way was cleared for reconciling the esu, emu, and practical systems. This was done by adopting the unit of length as the meter, rather than the centimeter, and the unit of mass as the kilogram rather than the gram, to give the present mks (meter-kilogram-second) system.

In the process, one other defect of both the esu and emu systems has been partially eliminated. Coulomb's law involves spherical geometry. Each charge experiences a force from the spherical electric field of the other charge. Since by definition $F = qE$, the field must be assigned unit strength with unit charges at unit distance when using Eq. (1-2-1) with $\epsilon = 1$. This leads to assigning a total flux $\Phi = 4\pi q$ to a charge q, and introduces the factor of 4π into a great many expressions involving rectangular goemetry and plane fields. The same situation exists in the emu system. In 1882 Oliver Heaviside pointed out that a "more rational" system of units would result if Eqs. (1-2-1) and (1-2-3) were written as

$$F = \frac{q_1 q_2}{4\pi\epsilon r^2} \tag{1-2-1r}$$

$$\frac{F}{l} = \frac{\mu I_1 I_2}{2\pi d} \tag{1-2-3r}$$

to take account of the spherical and cylindrical geometry respectively. The r in the equation numbers designates the *rational* form. Had this been done originally and ϵ and μ assigned the value unity for a vacuum, the esu and emu units of charge would have been $\sqrt{4\pi}$ times as great. Heaviside and Hendrik A. Lorentz proposed such a system and used it in their works. Modern practice is to write the equations in the above form,

use the esu, emu, or practical units, and subsume a compensating factor of 4π into ϵ and μ. This is the origin of the 4π's in the mks values of

$$\epsilon_0 = \frac{10^{11}}{4\pi c^2} \approx \frac{10^{-9}}{36\pi} \tag{1-2-5}$$

$$\mu_0 = 4\pi \times 10^{-7} \tag{1-2-6}$$

for the permittivity and permeability of free space; the remaining powers of 10 and c (the velocity of light) arise from the ratios of the esu and emu units to practical units. As a result of the rationalization, the factors of 4π are missing from planar field equations written in mks units, but are contained in μ and ϵ, which must always be included.

The assignment of dimensions to ϵ and μ in effect defines a new physical dimension which can be taken as the ampere or as the coulomb, and gets rid of the bizarre, irrational (and inconsistent) dimensions assigned to charge in the old esu and emu systems. For example, by virtue of Eq. (1-2-1) or (1-2-1r), ϵ in the mks system has the units

$$\frac{\text{coulomb}^2}{(\text{newton})\ (\text{meter})^2} = \frac{Q^2T^2}{ML^3} \tag{1-2-7}$$

We have no reason to believe that charge can be constructed out of mass, length, and time, and every reason to believe that charge is a qualitatively distinct physical quantity meriting a dimension of its own. Nevertheless, the introduction of this fourth unit into the absolute system of units caused long and often acrimonious debate. The defenders of the classical system of the three fundamental dimensions of mass, length, and time seem not to have been bothered by the two different dimensionalities of charge in the esu and emu systems. Further, they seemed unaware that even three fundamental dimensions are unnecessary. If γ in Eq. (1-2-4) is taken as dimensionless and equal to unity as was done for ϵ and μ in Eqs. (1-2-1) and (1-2-3), a new unit of mass is obtained having the dimensions L^3T^{-2}. Had this been done by Newton, then classicists who followed him might have objected violently to introducing a *third* "fundamental" dimension for mass.

As a matter of fact, even length and time can be eliminated by also defining, say, Planck's constant and the velocity of light as dimensionless and equal to unity. One then has a system of units in which all the variables in physical equations are pure numerics, and the unit size is set by the requirement that certain prescribed physical measurements produce identities. Such a system would obviate the need for including many constants of proportionality, but at the same time would greatly obscure the qualitative difference of the variables in all equations and prevent the use of dimensional analysis in checking calculations.

Since the number of fundamental dimensions in a system of units is thus somewhat arbitrary, there is certainly nothing magic about the number three and no good physical reason to exclude charge as a fundamental dimension.

It should be understood that the size of a unit can be chosen independent of its dimensionality. We are free to pick any convenient size for just the reason that it *is* convenient, i.e., the quantities we need to express involve neither inconveniently (and almost meaninglessly) large or small multiples of our unit. The standard prefixes in three order-of-magnitude steps help a great deal in this problem. In fact, the appeal of the metric system resides not in the length of the meter, nor the mass of the kilogram, but rather in the decimal subdivisions and multiples of these units.

The question arises: Is there a "natural" system of units? That is, can we choose the size of our units of mass, length, time, and charge so as to cause many or all of the physical constants to become integers or rational numbers or, at the very least, numbers that have some physical significance in terms of allowed degrees of freedom or the like? At present the answer to these questions appears to be no. We can, for example, define the mass and charge of the electron, the velocity of light, and Planck's constant as unity. But having done this, no further simplifications occur. No *other* physical constants assume integer values. Indeed there is one physical constant, the fine-structure constant, which combines the velocity of light, Planck's constant, the electronic charge, and the permittivity of vacuum in the equation

$$\alpha = \frac{\mu_0 c^2}{4\pi} \frac{e^2}{hc} = \frac{1}{137.03602\ldots}$$

The point is that α is dimensionless, so its value does not depend on our system of units. The number 137.03602 . . . is a natural physical number in the same sense that π or ε are natural mathematical numbers. Sir Arthur Stanley Eddington believed that α^{-1} had the integer value 137 and offered several hypotheses for why this might be true. More accurate modern measurements have disproved his premise. Until future study reveals the existence of a natural system of units, there appears to be no reason to abandon the present mks system.

1-3 Standards of Units of Measurement

In order to make accurate measurements in different places that are intercomparable, accurate standards are needed. The early standards were all prototype standards: physical objects that defined the unit as one of their physical properties. The standard kilogram and the standard

However, $R_{ss}(\tau) = \frac{1}{2}A^2 \cos \omega\tau$, and for large τ, $R_{nn}(\tau) \to 0$. Thus,

$$R_{zz}(\tau) = \tfrac{1}{2}A^2 \cos \omega\tau \tag{4-5-10}$$

for large τ.

We see therefore (Fig. 4-21) that by observing the autocorrelation function for large τ, it is possible to determine both the amplitude A and the frequency ω of the sinusoid.

APPENDIX 4-A

Determination of Impulse Response by Crosscorrelation

$$\begin{aligned} R_{xy}(\tau) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t-\tau)y(t)\,dt \\ &= \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t-\tau) \int_{-\infty}^{+\infty} h(u)x(t-u)\,du\,dt \end{aligned} \tag{4-A-1}$$

Interchange order of integration

$$R_{xy}(\tau) = \int_{-\infty}^{+\infty} h(u) \left[\lim_{T\to\infty} \frac{1}{T} \int_0^T x(t-\tau)x(t-u)\,dt \right] du \tag{4-A-2}$$

The second integral is seen to be the autocorrelation of $x(t)$ with argument $(\tau - u)$

$$R_{xy}(\tau) = \int_{-\infty}^{+\infty} h(u)R_{xx}(\tau - u)\,du \tag{4-A-3}$$

Now, if we choose the bandwidth of the noise to be much greater than the system passband, then in Eq. (4-A-3), $h(u)$ will be a relatively slowly changing function in comparison with $R_{xx}(\tau - u)$. The term $h(u)$will be almost constant over the small range of values of u around $u = \tau$ for which $R_{xx}(\tau - u)$ has significant values. The integral then becomes

$$R_{xy}(\tau) = h(\tau) \int_{-\infty}^{+\infty} R_{xx}(\tau - u)\,du \tag{4-A-4}$$

Comparing this with Eq. (4-1-21), we see that the integral term gives us the power spectral density for $f = 0$; hence

$$R_{xy}(\tau) = \tfrac{1}{2}h(\tau)G_x(0) \tag{4-A-5}$$

where $G_x(f)$ is the physically realizable, one-sided power density spectrum, and $G_x(0)$ is the value of this function in (volts)^2 per hertz, at very low frequencies.

APPENDIX 4-B
Effect of Contaminating Noise at Output of System

Crosscorrelation between the input $x(t)$ and the observable output $z(t)$,

$$R_{xz}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T x(t-\tau)z(t)$$

$$= \lim_{T\to\infty} \frac{1}{T}\left[\int_0^T x(t-\tau)y(t)\,dt + \int_0^T x(t-\tau)n(t)\,dt\right]$$

$$R_{xz}(\tau) = \int_{-\infty}^{\infty} h(u)R_{xx}(\tau-u)\,du + R_{xn}(\tau) \qquad (4\text{-B-1})$$

If $x(t)$ is white noise,

$$R_{xz}(\tau) = \tfrac{1}{2}G_x(0)h(\tau) + R_{xn}(\tau) \qquad (4\text{-B-2})$$

If background noise $n(t)$ is uncorrelated with test signal $x(t)$, then $R_{xn}(\tau) = 0$ for all τ. Hence,

$$R_{xz}(\tau) = \tfrac{1}{2}G_x(0)h(\tau) \qquad (4\text{-B-3})$$

That is, the result is unaffected by the presence of an uncorrelated background disturbance.

APPENDIX 4-C
Effect of Background Disturbances at Input to a System

The term $x(t)$ is a background disturbance and $s(t)$ is an artificially injected test signal.

$$R_{sy}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T s(t-\tau)y(t)\,dt \qquad (4\text{-C-1})$$

$$y(t) = \text{response due to } x(t) + \text{response due to } s(t)$$

$$= \int_{-\infty}^{\infty} h(u)x(t-u)\,du + \int_{-\infty}^{\infty} h(v)s(t-v)\,dv \qquad (4\text{-C-2})$$

Making the substitution for $y(t)$ and using a similar argument to that used in Appendix 4-A, we have

$$R_{sy}(\tau) = \int_{-\infty}^{+\infty} h(u)R_{sx}(\tau-u)\,du + \int_{-\infty}^{+\infty} h(v)R_{ss}(\tau-v)\,dv \qquad (4\text{-C-3})$$

If $s(t)$ and $x(t)$ are uncorrelated, then $R_{sx}(\tau)$ is zero for all τ, and if $s(t)$ is white noise, the second term is simplified. We then have

$$R_{sy}(\tau) = \tfrac{1}{2}G_s(0)h(\tau) \qquad (4\text{-C-4})$$

Crosscorrelation between the output and an injected white-noise test signal yields the system impulse response, regardless of the presence of an uncorrelated disturbance at the input.

APPENDIX 4-D
Statistical Errors in the Detection of a Sine Wave in Noise by Crosscorrelation

When the frequency of the sine wave is known, its amplitude and phase can be measured by crosscorrelating with a clean reference sine wave, even in the presence of large amounts of background noise. Statistical errors are involved when averages are taken over short intervals, and these may be analyzed as follows:

Let the noisy signal $x(t)$ consist of a clean sine wave $s(t)$ hidden in a background of band-limited gaussian noise $n(t)$ having a flat spectrum up to B Hz. The frequency f_c of the sine wave is less than the maximum noise frequency. We can write

$$s(t) = A \sin 2\pi f_c t \tag{4-D-1}$$
$$x(t) = s(t) + n(t) \tag{4-D-2}$$

where $n(t)$ is gaussian noise with spectrum flat to B Hz, $B > f_c$. The operations in crosscorrelation are indicated in Fig. 4-D-1. The statistical errors can be

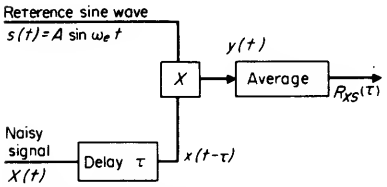


FIG 4-D-1 Diagram of the operations in crosscorrelation.

estimated by examining the statistics of the signal $y(t)$ at the output of the multiplier,

$$y(t) = s(t)[s(t - \tau) + n(t - \tau)] \tag{4-D-3}$$
$$y(t) = A^2 \sin 2\pi f_c t \sin 2\pi f_c (t - \tau) + n(t - \tau) A \sin 2\pi f_c t \tag{4-D-4}$$

If we take the average of $y(t)$ over an interval of T sec, there will be three components:

A mean level equal to the average of product

$$s(t)s(t - \tau) = \frac{A^2}{2} \cos 2\pi f_c \tau$$

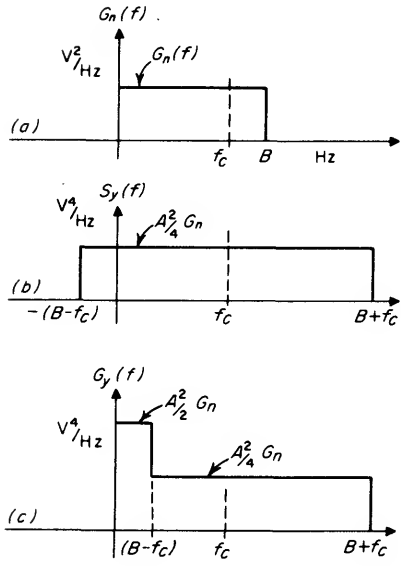


FIG 4-D-2 Power density spectra of $n(t)$ and $y(t)$.

A ripple component due to $s(t)s(t-\tau)$ at frequency $2f_c$. This will be completely smoothed out by averaging time T.

$$T \gg \frac{1}{f_c}$$

A component due to modulation products between noise $n(t)$ and sine wave $s(t)$. Noise in the band around the frequency f_c contributes to the variance in the estimate of $R_{sx}(\tau)$.

Only the first term gives rise to a dc component, and this is equal to the true value of the correlation function $R_{sx}(\tau)$ which we are trying to measure.

The problem is to evaluate the variance on the estimate of this mean value caused by the second and third terms, but since the second term is negligible, we shall concentrate on the third term. By using Eq. (4-2-5), the variance of the result can be written

$$\text{Var}\,[\hat{R}_{sx}(\tau)] = \text{var}\,(\hat{\mu}_y) = \frac{\sigma_{yn}^2}{2BT} \tag{4-D-5}$$

where σ_{yn} is the rms value of the random-noise component of $y(t)$ and B is its equivalent bandwidth, and this can be written in terms of the low-frequency power spectral density

$$\text{Var}\,[\hat{R}_{sx}(\tau)] = \text{var}\,(\hat{\mu}_y) = \frac{G_y(0)}{2T} \tag{4-D-6}$$

The magnitude of $G_y(0)$ can be deduced from a consideration of the power density spectra of $n(t)$ and $y(t)$. These are indicated in Fig. 4-D-2.

We can now compute the variance in the estimate of the mean value of $y(t)$.

$$\text{Var } (\hat{\mu}_y) = \frac{A^2 G_n}{4T} \tag{4-D-7}$$

where G_n is the power density of the original noise at frequencies in the region of the carrier frequency f_c.

$$\text{Var } [\hat{R}_{sx}(\tau)] = \frac{A^2 G_n}{4T} \tag{4-D-8}$$

The normalized error

$$\epsilon = \frac{\text{SD}[\hat{R}_{sx}(\tau)]}{R_{sx}(\tau)} = \sqrt{\frac{G_n}{A^2 T}} \tag{4-D-9}$$

CITED REFERENCES

1. Bendat, J. S., and A. G. Piersol: "Measurement and Analysis of Random Data," chap. 3, John Wiley & Sons, Inc., New York, 1966.
2. *Ibid.*, chap. 9.
3. *Ibid.*, chaps. 5 and 6.
4. *Ibid.*, p. 243.
5. *Ibid.*, chap. 6.
6. Golomb, Solomon W.: "Shift Register Sequences," Holden-Day, Inc., Publisher, San Francisco, 1967.
7. Anderson, G. C., B. W. Finnie, and G. T. Roberts: Pseudo-random and Random Test Signals, *Hewlett-Packard J.*, September, 1967.
8. Peterson, W. W.: "Error Correcting Codes," M.I.T. Press, Cambridge, Mass., 1961.
9. Rice, S. O.: Mathematical Analysis of Random Noise, in N. Wax, ed., "Selected Papers on Noise and Stochastic Processes," Dover Publications, Inc., New York, 1954.

CHAPTER FIVE

SIGNAL ANALYSIS BY DIGITAL TECHNIQUES

Ronald W. Potter

Hewlett-Packard Company,
Santa Clara, California

The theoretical aspects of signal analysis have been covered extensively in numerous texts and papers. However, the implementation of many of these operations on real data has not been studied so thoroughly. There are many restrictions imposed by "nature" that have a profound effect on the results of real-data analysis. For example, one restriction is the requirement of a finite observation interval. This chapter will discuss implications of these restrictions in the analysis of both random and coherent data. Data are assumed to be in digital form to facilitate the various operations that are normally required in signal analysis. The continual increase in availability and types of computer facilities and the decrease in cost are accelerating the development of digital instrumentation for signal processing.

In the context of this chapter, a signal is simply a function of a single real variable, usually time. The signal may be random, coherent, or mixed. It is generally desired to extract information from this signal in

some recognizable form. The various techniques that are available for this purpose comprise the tools for data analysis. The basic theory is assumed to be familiar and hence will be reviewed rather briefly.

It must be recognized that all signal analysis is done with a finite amount of data. Also, the data are available over some effective time interval, and there is a limit to the time resolution that can be obtained. The quotient of these two numbers (interval/resolution) is the number of data points available. Similar limits on range and resolution apply to the ordinate of each data point. The various analysis techniques may reduce or alter the data, but obviously cannot provide new information. Furthermore, the information content may be quite small in comparison with the amount of data available (in random noise, for example). Thus, it is necessary to recognize the uncertainty or statistical variation of each ordinate value, and it may take vast amounts of data to yield information with a small degree of uncertainty.

The Fourier transform and various statistical operations (such as averaging) constitute the basic tools that will be used. It is very convenient to have data in digital form for the implementation of these operations because arithmetic can be performed with arbitrary accuracy and data may be stored indefinitely. The digital approach also makes the various resolution and range limitations more explicit and helps to emphasize the finite amount of available information.

In order to place signal analysis in its proper context, consider some practical applications before delving into the theoretical details. Consider a vibrating structure such as an engine mount, a space vehicle, or a bridge, dam, or building. The designer must know the natural resonant frequencies and the sharpness (or Q) of these resonances. These quantities are usually very difficult to calculate and so must be measured. The excitation source may be a coherent signal such as a rotating eccentric weight attached to the face of a dam, or it may be a single transient such as a hammer blow. In some applications the source may be random noise such as turbulent air flow or a shake table driven from a random-noise generator. In many of these applications the source of excitation cannot be controlled by the user, which makes a study of the source characteristics another very important application. There may be many sources that contribute to the response at a particular point in a system. For example, how much of the vibration felt by an automobile passenger is caused by road roughness and how much originates in the motor or drive train? The use of crosscorrelation or cross-spectrum techniques allows the various sources and their propagation paths to be studied.

It may be desirable to test a control system with a prescribed driving function such as a pulse or random noise. The Fourier transform of the output waveform can be divided by the transform of the input to obtain

the system transfer function. Again, it is possible to simulate different transfer functions and then to evaluate the performance with different excitation signals.

In many situations the transfer function of a particular device or system causes a "smearing" of the true waveform and hence introduces a resolution limitation. In this case, a deconvolution or inverse-filtering technique can often be used to recover some of this "lost" resolution. Thus, the resolution of devices such as the gas chromatograph, the nuclear scintillation detector, and the sampling oscilloscope can be improved by using this technique. A similar scheme can be used to improve seismographic records that are obtained during exploration for oil or gas deposits.

The power spectrum "signature" of a device can be used for identification purposes. For example, underwater noises can be used to locate schools of fish or to identify ships and submarines.

Signal averaging has been used with considerable success in studying brain waves, heart waveforms, and pulsar signals from space.

5-1 *Fourier Transform—Review of Basic Theory* [1, 2, 3,]

Definition of continuous transform:

$$F(s) = \int_{-\infty}^{\infty} f(t)e^{-i2\pi st}\,dt \qquad \text{direct transform} \tag{5-1-1}$$

$$f(t) = \int_{-\infty}^{\infty} F(s)e^{i2\pi st}\,ds \qquad \text{inverse transform} \tag{5-1-2}$$

It is useful to think of t as time and s as frequency. Note that t and s have reciprocal dimensions. $F(s)$ has the dimensions of $tf(t)$ or $s^{-1}f(t)$. Thus, if $f(t)$ is in volts, $F(s)$ will be in volts per hertz.

Denote transform pairs with a double arrow. The shifting and scaling relation is given by

$$f\left(\frac{t-a}{b}\right) \leftrightarrow be^{-i2\pi as}F(bs) \tag{5-1-3}$$

where $f(t) \leftrightarrow F(s)$.

Contraction of the t axis by a factor b results in expansion of the s axis by b, but also introduces a multiplier of b which preserves the area in the s domain. Displacement of the t axis by a results in a helical twist about the s axis in the transform domain. The pitch of the helix is $-2\pi a$ rad/Hz.

Convolution is defined as follows:

$$h(t) = \int_{-\infty}^{\infty} f(\lambda)g(t-\lambda)\,d\lambda \tag{5-1-4}$$

Let $f(t) \leftrightarrow F(s)$, $g(t) \leftrightarrow G(s)$, and $h(t) \leftrightarrow H(s)$. Then $H(s) = F(s)G(s)$. Thus convolution in one domain corresponds to multiplication in the

other domain. It is very useful to develop a "feel" for the convolution process. As indicated in the definition, $g(t)$ is replaced by $g(-t)$ and multiplied by $f(t)$ for various displacements. The value of $h(t)$ is simply the area under the product curve for a particular displacement. Thus it should be obvious that a rectangle convolved with itself is a triangle. Any waveform convolved with a delta function is simply a replication of the original waveform.

From the defining equations for the transform the following relations can be obtained:

$$F(0) = \int_{-\infty}^{\infty} f(t)\,dt \qquad \text{and} \qquad f(0) = \int_{-\infty}^{\infty} F(s)\,ds \tag{5-1-5}$$

Thus, the area under $f(t)$ is simply the value of its transform at the origin. From this it can be deduced that the convolution operation multiplies areas. The area under $h(t)$ is the product of the areas under $f(t)$ and $g(t)$.

Integration and differentiation operations are

$$D^{(n)}F(s) \leftrightarrow (-i2\pi t)^n f(t) \qquad D^{(n)}f(t) \leftrightarrow (i2\pi s)^n F(s) \tag{5-1-6}$$

$D^{(n)}$ is the nth derivative operator. Integration is obtained when n is negative. Integration constants must usually be added.

By combining the various relationships previously defined with the commonly known transform pairs shown in Table 5-1, it is often fairly simple to deduce the transform of a particular function. The transform of a real function always produces a real part with even symmetry and an imaginary part with odd symmetry.

In practice, however, there are two reasons that the Fourier transform as defined above cannot be calculated:

1. Integration intervals must be finite.
2. Abscissa resolution is finite.

TABLE 5-1 List of Common Transform Pairs

$F(s)$	$f(t)$	
1	$\delta(t)$	Unit delta function
$\mathrm{III}(s)$	$\mathrm{III}(t)$	Infinite train of unit delta functions with unit spacing
$\frac{\sin \pi s}{\pi s}$	$\sqcap(t)$	Rectangle with unit width, height, and area
$e^{-\pi s^2}$	$e^{-\pi t^2}$	Gaussian with unit height and area
$\frac{1}{2}\left[\delta(s) - \frac{i}{\pi s}\right]$	$U(t)$	Unit step
$\frac{i}{2}[\delta(s+1) - \delta(s-1)]$	$\sin 2\pi t$	
$\frac{1}{2}[\delta(s+1) + \delta(s-1)]$	$\cos 2\pi t$	

As derived below, finite integration in one domain implies finite abscissa resolution in the other domain. Assume that t is restricted to the range $0 \leq t < T$. Since there is no available information outside of this range, an uncertainty has been introduced. This restricted function can be

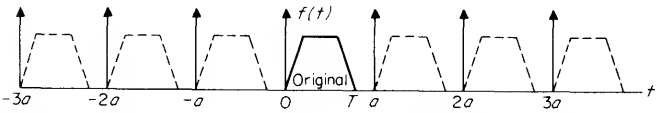


FIG 5-1 A function convolved with an infinite train of delta functions.

convolved with an infinite train of unit delta functions $a^{-1}\mathrm{III}(t/a)$ spaced a units apart without affecting the interval $0 \leq t < T$, provided $a \geq T$ (see Fig. 5-1). Thus $F(s)$ can be multiplied by $\mathrm{III}(sT)$ without altering the original information. These samples are spaced $\Delta s = 1/T$ apart. A similar argument will show that finite time resolution Δt dictates a band-limited frequency function not exceeding $\pm 1/2\Delta t$. These restrictions imply a finite number of data points, $N = T/\Delta t = 1/(\Delta s\ \Delta t)$.

Introducing the functions $\mathrm{III}(s/\Delta s)$ and $\mathrm{III}(t/\Delta t)$ into the Fourier transform definition causes the variables t and s to take on discrete values ($t = n\ \Delta t$ and $s = m\ \Delta s$) and converts the integrations over infinite intervals into summations over finite intervals. The resulting discrete Fourier transform definitions are given below, with $N\ \Delta t\ \Delta s = 1$.

$$F(m\ \Delta s) = \Delta t \sum_{n=0}^{N-1} f(n\ \Delta t) e^{-i2\pi mn/N} \tag{5-1-7}$$

$$f(n\ \Delta t) = \Delta s \sum_{m=0}^{N-1} F(m\ \Delta s) e^{i2\pi mn/N} \tag{5-1-8}$$

Now the Fourier transform simply becomes the product of a vector $\mathbf{f}(n\ \Delta t)$ and a matrix $e^{-i2\pi mn/N}$ to yield a vector $\mathbf{F}(m\ \Delta s)$ in the transform domain. A conjugate matrix is used for inversion. This is the classical way of calculating the transform, but it requires a very large number of arithmetic operations. In contrast, the Cooley-Tukey algorithm [4, 5] is much more efficient and will be described very briefly.†

If N is factorable, then m and n can be written as the sum of several indices, each corresponding to a factor. Thus, the single large summation

† Though the following few paragraphs will not enable the reader to use the algorithm in computations, it will give him a brief introduction to the procedure. Consult the references [4, 5] for detailed instructions.

is decomposed into several small summations. Similarly, the exponential matrix is broken into the product of several smaller matrices. The most efficient factorization occurs when N is an integer power of 2. Define a new set of indices as follows:

$$n = \sum_{r=0}^{R-1} k_r 2^r \tag{5-1-9}$$

where $N = 2^R$

$$m = \sum_{r=0}^{R-1} j_r 2^r \tag{5-1-10}$$

where j_r and k_r are binary variables.

The product mn contains terms that are multiples of N. These terms simply introduce a factor of unity and hence may be ignored. The resulting expression for mn/N is given below.

$$\frac{mn}{N} = \sum_{s=0}^{R-1} \sum_{r=0}^{R-s-1} j_r k_s 2^{s+r-R} \tag{5-1-11}$$

The discrete Fourier transform may now be written as follows:

$$F(j_0,j_1, \ldots ,j_{R-1}) = \Delta t \sum_{k_0=0}^{1} \sum_{k_1=0}^{1} \cdots \sum_{k_{R-1}=0}^{1} f(k_0,k_1, \ldots ,k_{R-1}) \prod_{s=0}^{R-1} \prod_{r=0}^{R-s-1} e^{-i2\pi Q_{jk}} \tag{5-1-12}$$

where

$$Q_{jk} = j_r k_s 2^{s+r-R}$$

A study of mn/N will show that each time k_r is summed out of the expression it is replaced by j_{R-r-1}. Thus the least significant indices of n are replaced by the most significant indices of m. After R summations, the k indices have been summed out and replaced by the j indices. Since the significance of the indices has been reversed, the transform data will not be ordered properly and thus must be reordered. A study of mn/N will show that after each summation there are additional multiplying factors that must be applied. Some of these must be applied before the next summation, while the application of other factors is optional. This algorithm essentially allows a one-dimensional transform to be calculated as an R-dimensional transform. However, the R dimensions are not independent since the original data points are ordered. The multiplying

factors after each summation are necessary to preserve the ordered relationship between the points.

The discrete Fourier transform is not the same as the continuous transform, and the resulting frequency functions may be considerably different. As indicated previously, sampling in the time domain with an interval Δt requires that the frequency spectrum be band limited to $\pm 1/2\Delta t$. Thus the minimum sampling rate (Nyquist rate) is twice the highest frequency present. If sampling is done slower than this, there will be an overlap of adjacent replicas of the true spectrum and "aliasing" will occur. In some cases this overlap is actually desired, although in most cases it is considered an *aliasing* error. Sampling in the time domain with an interval Δt actually causes a superposition of all frequency intervals spaced $1/\Delta t$ apart. Thus all frequencies above the Nyquist rate will be *heterodyned* into the base band. These "image" frequencies cannot be distinguished from the true base-band spectrum.

Another very important anomaly introduced by the discrete transform is the finite time window and resulting frequency line shape. A rectangular time window of width T produces a $\sin \pi s T/\pi s$ line shape in the frequency domain. Thus the true spectrum is convolved (smeared) by this line shape. This convolution is effected before quantization along the frequency axis. Thus, if the original data are periodic such that there is an integer number of periods in the time window, the line shape does not appear because frequency samples occur at zero crossings. However, if there is an odd number of half cycles in the time window, frequency samples occur at the peaks of the $\sin \pi s T/\pi s$ line shape and a hyperbolic line shape $1/\pi s$ occurs. Thus a single frequency waveform may transform into a broad band of frequencies. This *leakage* from one frequency into a band of adjacent frequencies is usually very undesirable. The cure is to reshape the time window to produce a line shape with smaller side lobes. This technique will broaden the main lobe and hence reduce frequency resolution, but it reduces the interference of one frequency with its neighbors. Various time windows and line shapes will be evaluated later.

References 1 through 5 are useful for the discussion to this point in the chapter.

5-2 *Statistics—Review of Basic Theory* [6, 7, 8]

A random variable is normally described by its pdf. Consider a random variable x and its probability density function $p(x)$. Then $\int_b^{b+\Delta x} p(x)\,dx$ is the probability of finding x between b and $b + \Delta x$. Obviously, the total area under $p(x)$ is unity. In a similar way, it is possible to describe multidimensional pdf's of several random variables.

The characteristic function of a random variable is simply the Fourier transform of its pdf.

$$F(s) = \int_{-\infty}^{\infty} p(x)e^{-i2\pi xs}\,dx \tag{5-2-1}$$

$$p(x) = \int_{-\infty}^{\infty} F(s)e^{i2\pi xs}\,ds \tag{5-2-2}$$

Note that $F(0) = \int_{-\infty}^{\infty} p(x)\,dx = 1$.

Differentiation of Eq. (5-2-1) n times and evaluation at $s = 0$ gives

$$F^{(n)}(0) = \int_{-\infty}^{\infty} (-i2\pi x)^n p(x)\,dx = (-i2\pi)^n \overline{x^n} \tag{5-2-3}$$

where $\overline{x^n}$ is the nth moment of $p(x)$

$$\overline{x^n} = \frac{F^{(n)}(0)}{(-i2\pi)^n} = \int_{-\infty}^{\infty} x^n p(x)\,dx \tag{5-2-4}$$

Thus, the moments of $p(x)$ can be easily calculated from the derivatives of the characteristic function. Similar equations apply for multidimensional distributions.

The cumulative probability distribution is defined as

$$P(x) = \int_{-\infty}^{x} p(\lambda)\,d\lambda \tag{5-2-5}$$

This expresses the probability that the random variable is less than some value x. The expression $P(x)$ is used to obtain limits within which the random variable will be found a prescribed percentage of the time.

By using the characteristic function, it is easy to show that the probability density of the sum of several independent random variables is simply the convolution of their respective probability densities. Define

$$x = \sum_{i=1}^{n} x_i$$

as the sum of n random variables. Then

$$p(x) = p_1(x_1) * p_2(x_2) * \cdots p_n(x_n)$$

where $p_i(x_i)$ is the probability density of the random variable x_i, and the asterisk denotes convolution.

A random process can be described simply as a random variable that changes with time. Thus, in addition to the description of a random variable, it is necessary to describe the dynamics of the process as a function of time. The study of a single random record has limited value; therefore, the concept of an ensemble of records of a random process must

be introduced. Consider an arbitrarily large number of random records being produced simultaneously from different but identical sources. These data may be recorded in a multichannel memory for further study. The ensemble properties may be studied for each *time* value, or the time properties may be studied for each *record* in the ensemble.

A stationary process is one in which the ensemble statistics (pdf and moments) are independent of time. Otherwise it is called a nonstationary random process. A stationary process is ergodic if the ensemble statistics are identical with the time statistics. This means that ensemble averages are the same as time averages. Most of the theoretical work in this field is based on stationary (and usually ergodic) processes. However, nonstationary processes very often occur in practice.

The dynamics of a stationary random process are described by a power density spectrum or its Fourier transform, the autocorrelation function. It should be emphasized that the power density spectrum is *not* related to the pdf, and one cannot be derived from the other. It is possible that the pdf will be different in each frequency interval of the power density spectrum. The power density spectrum is an ensemble average of the power density of each time record. If the process is ergodic, an average of power densities calculated for different times may be used in place of an ensemble average.

5-3 *Signal Analysis*

The extraction of meaningful information from an arbitrary signal is more of an art than a science in many cases. If nothing is known a priori about the signal, then several measurement techniques must generally be tried and the properties of these measurement processes must be known in order to interpret the results correctly. The questions arise: How many power density spectra must be averaged to allow a prescribed accuracy in the detection or measurement of a coherent signal buried in noise? What is the frequency-domain equivalent of linear signal averaging? What are the problems encountered in inverse filtering or deconvolution? These questions and others of a related nature will be discussed, along with the various measurement techniques that have been found to be useful.

Histograms. One of the simplest statistical measures is the amplitude probability density, or amplitude histogram. The general technique is to digitize the instantaneous amplitude of a signal and then to increment the count in a memory cell whose address is given by this signal amplitude. The statistical significance of the number in each memory cell is dependent on the magnitude of the number. This relation will be derived below.

It is important to note that the differential linearity of the analog-to-digital converter is very critical in histogram applications. To understand this, denote the voltage resolution of the analog-to-digital converter by ΔE. Ideally each ΔE would be identical, but in practice there is some variation in ΔE from one voltage level to another. A typical probability density is shown in Fig. 5-2.

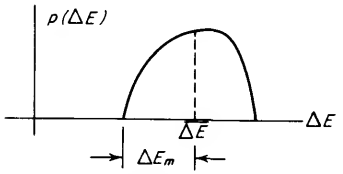


FIG 5-2 Probability density function of analog-to-digital converter sampling error.

Define ΔE_m (Fig. 5-2) as the maximum deviation from $\overline{\Delta E}$ (mean value of ΔE). Then the differential linearity is $\eta = \Delta E_m/\overline{\Delta E}$. In histogram measurements the amplitude in each channel is directly proportional to the channel width ΔE, and thus the histogram plot will have an amplitude distribution that is the same as shown in Fig. 5-2. For example, 1 percent differential linearity will produce 1 percent amplitude variation in the histogram distribution.

Assume that N memory channels are used to accumulate an amplitude histogram. Define ϵ as the probability of finding the input voltage in some particular memory channel. The probability density $p_1(m)$ for m counts in a particular channel after only one count has been recorded is shown in Fig. 5-3. (The *probability* is the area of the delta function.)

FIG 5-3 Probability density after one count (see text).

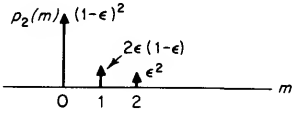


FIG 5-4 Probability density after two counts (see text).

The probability density for m counts in a particular channel after a total of two counts have been recorded is simply the convolution of $p_1(m)$ with itself (Fig. 5-4). By generalizing this concept, it can be seen that the probability for m counts in a particular channel after a total of n

counts have been recorded is given by

$$p_n(m) = \binom{n}{m} \epsilon^m (1 - \epsilon)^{n-m} \tag{5-3-1}$$

where

$$\binom{n}{m} = \frac{n!}{m!(n - m)!}$$

are the binomial coefficients. This equation represents a binomial distribution. For $n\epsilon(1 - \epsilon) \gg 1$, the above expression becomes gaussian:

$$p_n(m) \approx \frac{1}{[2\pi n\epsilon(1 - \epsilon)]^{1/2}} e^{-(m-n\epsilon)^2/[2n\epsilon(1-\epsilon)]} \tag{5-3-2}$$

This equivalence is called the *DeMoivre-Laplace theorem* [9].

The mean is $\bar{m} = n\epsilon$ and the standard deviation is $\sigma = \sqrt{n\epsilon(1 - \epsilon)}$. The quantity $\sigma/m = [(1 - \epsilon)/n\epsilon]^{1/2}$ is a good measure of the statistical uncertainty. Note that this uncertainty depends on ϵ as well as n. For $\epsilon = 1$, all counts go into one channel and $\sigma = 0$. For $\epsilon < 1$, the uncertainty increases for a given value of n. The value of ϵ may not be known a priori, but it will become apparent as the histogram is formed.

Histograms are very useful to detect the presence of a coherent signal buried in noise, particularly when the noise has a gaussian or other known distribution. Amplitude histograms of coherent signals can often be recognized. Define a coherent and periodic function $z = f(t)$. It can readily be shown that the probability density of z is given by

$$p(z) = \frac{1}{Tf'(t)} \qquad \text{for } -\frac{T}{2} \leq t \leq \frac{T}{2} \tag{5-3-3}$$

where T is the period of $f(t)$. Thus the probability density of a coherent signal is inversely proportional to its slope. For example, define $z = f(t) = \sin t$. Then $f'(t) = \cos t = (1 - z^2)^{1/2}$, $T = 2\pi$, and $-\pi \leq t \leq \pi$. So, $p(z) = 1/[\pi(1 - z^2)^{1/2}]$ for $-1 \leq z \leq 1$. The range of z is covered twice as t ranges between $\pm\pi$, and so $p(z)$ includes this additional factor of 2.

Signal Averaging. For ergodic random processes it is possible to obtain an ensemble average by selecting records in time sequence. If the process has some coherent signal, and if sampling is done in synchronism with this coherent part, then the ensemble average will increase the signal-to-noise ratio. There are two aspects of this signal-averaging process to be considered. One aspect is concerned with the statistical variation in each memory channel of the sampled signal, and the other aspect is concerned with the dynamics or time-frequency behavior of the process. Consider first the statistical variation in each channel. All samples for a particular

memory channel are drawn from the same population, whose probability density can be described by Fig. 5-5. There exists some arbitrary probability density function $p(x)$ with a mean $\bar{x}$ that represents the coherent signal value for that particular channel, and a distribution about that

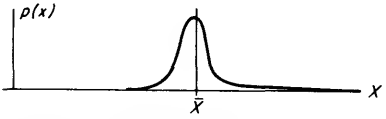


FIG 5-5 Probability density in one memory channel of signal averager.

mean which represents random noise. Let σ^2 be the variance of $p(x)$. The averaging process consists of summing n random variables from the above population (for each memory channel) and then dividing by n to obtain an average. If the samples are independent, the probability density of the sum of these n random variables is simply $p(x)$ convolved together n times. This will produce a new mean $n\bar{x}$ and a new variance $n\sigma^2$. The function $p(x)$ will also tend to become gaussian for large n (central-limit theorem). Division by n will restore the mean to $\bar{x}$, but the variance is reduced to σ^2/n and the standard deviation is $\sigma/(n)^{1/2}$. Thus the averaging process has reduced the noise voltage by $\sqrt{n}$, provided all samples are independent.

The temporal dynamics of the signal-averaging process are best described by deriving the *equivalent filter* of the process in the frequency domain. Consider an arbitrary signal $f(t)$ with a periodic component of period τ. Records of length T will be sampled with a time window $g(t)$ and averaged into a memory. Successive records are taken in synchronism with the periodic component of $f(t)$ and are spaced $k\tau$ apart, where k is some positive integer (often unity). The resulting average of N time records can be represented by

$$h(t) = \frac{1}{N} \sum_{n=0}^{N-1} f(t')g(t') \tag{5-3-4}$$

where $t' = t - nk\tau$ and $g(t) = 0$ outside the interval $0 \leq t \leq T$. Define $H(s) \leftrightarrow h(t)$, $F(s) \leftrightarrow f(t)$, and $G(s) \leftrightarrow g(t)$ as Fourier transform pairs. Then $H(s)$ becomes

$$\begin{aligned} H(s) &= \frac{1}{N} \sum_{n=0}^{N-1} [F(s) * G(s)]e^{-i2\pi nk\tau s} \\ &= \frac{1}{N} [F(s) * G(s)] \sum_{n=0}^{N-1} e^{-i2\pi nk\tau s} \end{aligned} \tag{5-3-5}$$

where the asterisk denotes the convolution operation. The equivalent filter for the signal-averaging operation is

$$\phi(s) = \frac{1}{N}\sum_{n=0}^{N-1} e^{-i2\pi nk\tau s} = \frac{1}{N}\frac{1 - e^{-i2\pi Nk\tau s}}{1 - e^{-i2\pi k\tau s}} \tag{5-3-6}$$

$$= \frac{\sin \pi Nk\tau s}{N \sin \pi k\tau s} e^{-i\pi(N-1)k\tau s}$$

$$|\phi(s)| = \frac{\sin \pi Nk\tau s}{N \sin \pi k\tau s} \tag{5-3-7}$$

The function $|\phi(s)|$ resembles a comb filter with major lobes at $s = m/k\tau$, where m is an integer (see Fig. 5-6). The quantity $1/k\tau$ is the rate at which the various records of data are acquired. Thus all harmonics of this data-acquisition rate are preserved in the signal averaging. There are minor lobes at $s = (m + \frac{1}{2})/Nk\tau$, with an envelope $(1/N)$ csc $\pi k\tau s$. Further, $|\phi(s)| = 1$ for $s = m/k\tau$ (major lobes), and $|\phi(s)| = 1/N$ at the midpoint between major lobes. Thus various frequencies are attenuated by different amounts depending on their relationship with the basic data-acquisition rate. There are $N - 1$ lobes, and the main lobe is twice as wide as the side lobes. Thus as more records are averaged, the main lobe narrows, more side lobes appear, and the attenuation between major lobes increases. The following relation can be readily derived:

$$\int_0^{1/k\tau} \phi(s)\phi^*(s)\, ds = \int_0^{1/k\tau} |\phi(s)|^2\, ds = \frac{1}{N} \tag{5-3-8}$$

Thus, the power of initially white noise is reduced by $1/N$ by the averaging process, although the noise is no longer white.

At the end of the chapter (Fig. 5-32) the effective transfer functions of a signal averager are shown, as determined by a digital computer. Starting with the top photograph, the value for N is progressively increased: 2, 4, 8, 16, 32.

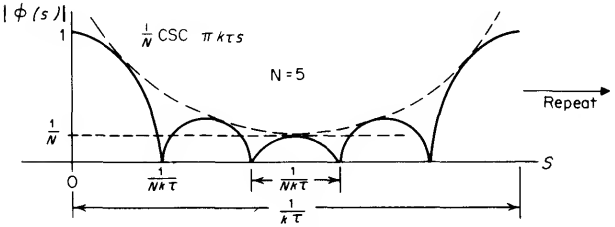


FIG 5-6 Equivalent filter for a signal averager with $N = 5$.

Digitizing Errors. A digression will be made at this point to consider the effects produced in converting an analog signal into a digital representation. Consider a periodic record of period T sampled at Δt intervals so that $T = n\,\Delta t$. Each sample is digitized in an analog-to-digital con-

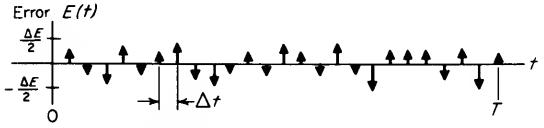


FIG 5-7 Error signal after sampling.

verter with a voltage resolution ΔE. Each sample may be in error by $\pm\Delta E/2$ or less, and Fig. 5-7 shows one possible appearance of the error signal after sampling. Assume that the error in each sample is random and independent of all other sample errors. The probability density

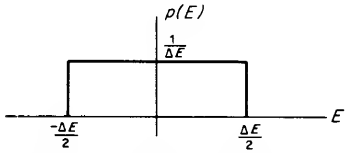


FIG 5-8 Probability density of sampling error.

(Fig. 5-8) of this random variable is rectangular. The variance of E is $(\Delta E)^2/12$ and represents the total noise power. The autocorrelation function is simply a train of delta functions spaced T apart, each having an area of $\Delta t\ (\Delta E)^2/12$. The power density spectrum (Fig. 5-9) is there-

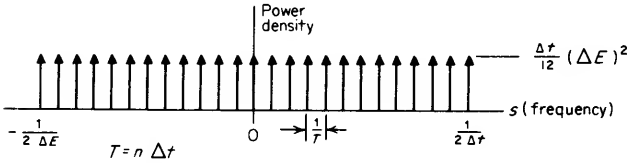


FIG 5-9 Power density spectrum of sampling error.

fore flat with frequency and has an amplitude of $\Delta t\ (\Delta E)^2/12$. There are n samples of the power density spaced $1/T$ apart, so that the total power is $(\Delta E)^2/12$. Suppose that an m-bit analog-to-digital converter is used so that a full-scale voltage would have a magnitude $E_m = \pm 2^{m-1}\,\Delta E$. A full-scale pulse of width Δt would produce a flat power density spectrum

of amplitude $\Delta t\, E_m^2 = \Delta t\, (\Delta E)^2\, 2^{(2m-2)}$. The power signal-to-noise ratio at each frequency is 3×2^{2m}. Thus for a 10-bit analog-to-digital converter the signal-to-noise ratio is 64.98 dB. If a full-scale dc signal is sampled, the power density spectrum would be a single sample point of amplitude $n\, \Delta t\, E_m^2$. Thus the power signal-to-noise ratio is improved by the number of sample points in the record. It must be emphasized that the above results are only valid if the errors at the various sample points are independent. When sampling coherent waveforms, this condition may not be met. The worst case for coherent sampling occurs when the error is constant over the record with an amplitude of $\Delta E/2$. The power spectrum is a single sample point of amplitude $n\, \Delta t\, (\Delta E)^2/4$. The worst power signal-to-noise ratio is $2^{2m}/n$ at dc.

There is a technique for reducing the noise caused by the digitizing process, provided it is possible to average many signal records. The scheme simply is to add a small amount of noise to the signal before sampling, and then average the result to reduce the noise. If the desired digitizing noise reduction is specified, the number of records that must be averaged is determined thereby, and the amount of noise that must be added to the original signal can be calculated. The algebra becomes rather tedious for portions of the discussion which follows, and so in some cases only the results will be shown. However, the general method of attack will be explained.

Assume an arbitrary nonrandom input signal $E(t)$, which is sampled at some particular time t_0 (see Fig. 5-10). Assume that the true voltage at t_0 is E_s. The pdf for this voltage (Fig. 5-11)is $\delta(E - E_s)$. Next assume random noise with probability density $p_0(E)$ is added to $E(t)$. The probability density of the sum of two independent random variables is the convolution of their individual probability densities. The resulting distribution is shown in Fig. 5-12. The mean of $p_0(E)$ is assumed to be

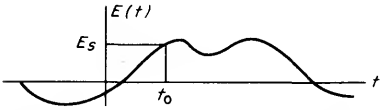


FIG 5-10 An arbitrary nonrandom input signal.

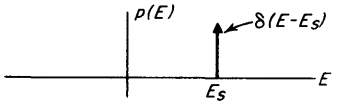


FIG 5-11 Probability density for voltage at t_0 in Fig. 5-10.

zero. In the sampling process the E axis is digitized into intervals of width ΔE (analog-to-digital converter resolution). The value assigned to the mth point is simply the integral of $p(E)$ over an interval of width ΔE centered on the mth point. Figure 5-13 should clarify this statement.

$$p(m\, \Delta E) = \delta(E - m\, \Delta E) \int_{(m-\frac{1}{2})\Delta E}^{(m+\frac{1}{2})\Delta E} p_0(\lambda - E_s)\, d\lambda \qquad (5\text{-}3\text{-}9)$$

The probability density after sampling then becomes

$$p_1(E) = \sum_{m=-\infty}^{\infty} p(m\,\Delta E)$$
$$= \sum_{m=-\infty}^{\infty} \delta(E - m\,\Delta E) \int_{-\infty}^{\infty} \sqcap\left(\frac{\lambda - m\,\Delta E}{\Delta E}\right) p_0(\lambda - E_s)\,d\lambda \tag{5-3-10}$$

where $\sqcap(E)$ is a rectangle of unit width and height centered at the origin. Note that $\sqcap(E) = \sqcap(-E)$. Thus the integral is represented by the con-

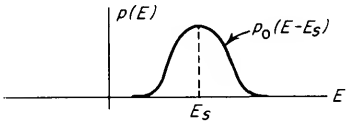


FIG 5-12 Probability density for $E(t)$ with random noise added.

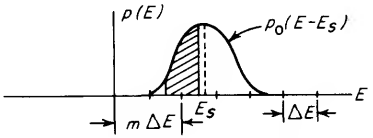


FIG 5-13 Digitized value (shaded area) of $p(E)$.

volution of $\sqcap(E/\Delta E)$ and $p_0(E - E_s)$. Define

$$\text{III}\left(\frac{E}{\Delta E}\right) = \Delta E \sum_{m=-\infty}^{\infty} \delta(E - m\,\Delta E)$$

Thus,

$$p_1(E) = \frac{1}{\Delta E}\left[\sqcap\left(\frac{E}{\Delta E}\right) * p_0(E - E_s)\right] \text{III}\left(\frac{E}{\Delta E}\right) \tag{5-3-11}$$

The characteristic function $\Pi_1(s)$ is given by

$$\Pi_1(s) = \text{III}(s\,\Delta E) * \left[\Pi_0(s)\,\frac{\sin \pi s\,\Delta E}{\pi s}\,e^{-i2\pi s E_s}\right] \tag{5-3-12}$$

where $\Pi_0(s) \leftrightarrow p_0(E)$ are Fourier transform pairs. Thus the continuous function within the brackets is reproduced about a train of delta functions spaced $1/\Delta E$ apart. Suppose that n records from the analog-to-digital converter are averaged together. The probability density of the sum of n independent random variables is the convolution of their individual densities. The characteristic function is therefore the product of the individual characteristic functions.

$$p_n(E) = p_1(E)^{*n} \tag{5-3-13}$$

where * denotes convolution

$$\Pi_n(s) = \Pi_1{}^n(s) \tag{5-3-14}$$

Writing Eq. (5-3-12) in a different way,

$$\Pi_1(s) = \sum_{m=-\infty}^{\infty} \frac{\sin \pi\ \Delta E\ z_m}{\pi\ \Delta E\ z_m} \Pi_0(z_m) e^{-i2\pi E_s z_m} \tag{5-3-15}$$

where $z_m = s - (m/\Delta E)$. Note that $\Pi_1(0) = \Pi_0(0) = 1$ and $\Pi_0^{(1)}(0) = 0$. Recall that the rth moment of E is given by

$$\overline{E^r} = \frac{\Pi^{(r)}(0)}{(-i2\pi)^r} \tag{5-3-16}$$

where $\Pi^{(r)}(s)$ is the rth derivative of $\Pi(s)$. Thus it is necessary to evaluate at least the first two derivatives of $\Pi_n(s)$ at $s = 0$.

$$\Pi_n^{(1)}(s) = n\Pi_1^{n-1}(s)\Pi_1^{(1)}(s) \tag{5-3-17}$$

$$\Pi_n^{(2)}(s) = n\Pi_1^{n-1}(s)\Pi_1^{(2)}(s) + n(n-1)\Pi_1^{n-2}(s)[\Pi_1^{(1)}(s)]^2 \tag{5-3-18}$$

$$\Pi_1^{(1)}(s) = \sum_{m=-\infty}^{\infty} \left\{ \frac{az_m \cos az_m - \sin az_m}{az_m^2} \Pi_0(z_m) + \frac{\sin az_m}{az_m} [\Pi_0^{(1)}(z_m) - i2\pi E_s \Pi_0(z_m)] \right\} e^{-i2\pi E_s z_m} \tag{5-3-19}$$

where $z_m = s - \dfrac{m}{\Delta E}$

$a = \pi\ \Delta E$

$$\Pi_1^{(1)}(0) = \sum_{m=-\infty}^{\infty} (-1)^m \frac{\Delta E}{m} \Pi_0\left(\frac{m}{\Delta E}\right) e^{-i2\pi m E_s/\Delta E} \tag{5-3-20}$$

for $m \neq 0$, and for $m = 0$,

$$\Pi_1^{(1)}(0) = \Pi_0^{(1)}(0) - i2\pi E_s = -i2\pi E_s \tag{5-3-21}$$

The second derivative is straightforward but involves considerable algebra. Only the result for $\Pi_1^{(2)}(0)$ will be given.

$$\Pi_1^{(2)}(0) = -(2\pi E_s)^2 - 2 \sum_{m=-\infty}^{\infty} (-1)^m \frac{\Delta E}{m} \left[\frac{\Delta E}{m} \Pi_0\left(\frac{m}{\Delta E}\right) + i\pi E_s \Pi_0\left(\frac{m}{\Delta E}\right) - \Pi_0^{(1)}\left(\frac{m}{\Delta E}\right) \right] e^{-i2\pi m E_s/\Delta E} \tag{5-3-22}$$

for $m \neq 0$.

$$\Pi_1^{(2)}(0) = -(2\pi)^2 E_s^2 - \tfrac{1}{3}(\pi\ \Delta E)^2 + \Pi_0^{(2)}(0) \tag{5-3-23}$$

for $m = 0$.

$$\Pi_n^{(1)}(0) = n\,\Pi_1^{(1)}(0) \tag{5-3-24}$$

$$\Pi_n^{(2)}(0) = n\Pi_1^{(2)}(0) + n(n-1)[\Pi_1^{(1)}(0)]^2 \tag{5-3-25}$$

The variance can be obtained by subtracting the square of the mean from the second moment. In the s domain this gives

$$\Pi_n^{(2)}(0) - [\Pi_n^{(1)}(0)]^2 = n\Pi_1^{(2)}(0) - n[\Pi^{(1)}(0)]^2 \tag{5-3-26}$$

Thus the variance after summing is simply n times the variance before summing. Notice that the moments before sampling are given by $m = 0$. Thus the errors in the moments introduced by sampling are given by the terms $m \neq 0$. The moments of E after sampling and averaging are

$$\bar{E} = E_s + \epsilon \tag{5-3-27}$$

$$\overline{E^2} - (\bar{E})^2 = \frac{1}{n}\left[\frac{1}{12}(\Delta E)^2 - \frac{1}{4\pi^2}\Pi_0^{(2)}(0)\right] + \delta^2 \tag{5-3-28}$$

$$\epsilon = \frac{i}{2\pi}\sum_{m=-\infty}^{\infty}(-1)^m\frac{\Delta E}{m}\Pi_0\left(\frac{m}{\Delta E}\right)e^{-i2\pi mE_s/\Delta E} \tag{5-3-29}$$

for $m \neq 0$.

$$\delta^2 = \frac{1}{2n\pi^2}\sum_{m=-\infty}^{\infty}(-1)^m\frac{\Delta E}{m}\left[\frac{\Delta E}{m}\Pi_0\left(\frac{m}{\Delta E}\right) - i\pi E_s\Pi_0\left(\frac{m}{\Delta E}\right) - \Pi_0^{(1)}\left(\frac{m}{\Delta E}\right)\right]e^{-i2\pi mE_s/\Delta E} + (\overline{\Delta E})^2 \tag{5-3-30}$$

for $m \neq 0$.

$$(\overline{\Delta E})^2 = \frac{1}{4n\pi^2}\left[\sum_{m=-\infty}^{\infty}(-1)^m\frac{\Delta E}{m}\Pi_0\left(\frac{m}{\Delta E}\right)e^{-i2\pi mE_s/\Delta E}\right]^2 \tag{5-3-31}$$

for $m \neq 0$.

The errors caused by sampling are ϵ and δ^2. The quantity $(\overline{\Delta E})^2 \approx 0$ for all practical purposes. Notice that these errors are represented in Fourier series form in the variable $E_s/\Delta E$. The errors depend on the precise location of E_s in the ΔE interval of analog-to-digital converter resolution.

For most practical situations, the repeated convolution of probability densities will eventually lead to a gaussian probability density (central-limit theorem) so that it is reasonable to assume a gaussian distribution for E. The moments have already been calculated; therefore, the distribution is given explicitly by

$$p_n(E) = \frac{1}{\Sigma(2\pi)^{1/2}}e^{-\frac{1}{2}(E-\bar{E})/\Sigma^2}\,\mathrm{III}\left(\frac{nE}{\Delta E}\right) \tag{5-3-32}$$

where $\Sigma^2 = \overline{E^2} - (\bar{E})^2$. The random variable E actually is a discrete variable with an interval $\Delta E/n$ (the averaging process has reduced the interval by n). However, $p_n(E)$ may be considered a continuous function for all practical purposes.

Define a confidence level A as follows:

$$A = \int_{E_s - E_0}^{E_s + E_0} p_n(E)\, dE \tag{5-3-33}$$

where E_0 is a confidence interval. This means that the probability of finding E in the interval $E_s - E_0 \leq E \leq E_s + E_0$ is A. Normally A is specified and E_0 is calculated. For the gaussian distribution given above, A can be written as

$$A = \frac{1}{2}\left[\operatorname{erf}\left(\frac{E_0 + \epsilon}{2^{1/2}\Sigma}\right) + \operatorname{erf}\left(\frac{E_0 - \epsilon}{2^{1/2}\Sigma}\right)\right] \tag{5-3-34}$$

where erf (z) is the error function. Given A, then E_0 can be calculated or determined from a table of the error function.

Consider a special case where the additive noise before sampling has a gaussian probability density with zero mean.

$$p_0(E) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-\frac{1}{2}(E/\sigma)^2} \tag{5-3-35}$$

where σ^2 is the variance and rms noise power. The characteristic function is

$$\Pi_0(s) = e^{-2(\sigma\pi s)^2} \tag{5-3-36}$$

The error in mean value caused by the digitizing process is

$$\epsilon = \frac{1}{\pi} \sum_{m=1}^{\infty} (-1)^m \frac{\Delta E}{m} e^{-2(m\pi\sigma/\Delta E)^2} \sin\left(2\pi m \frac{E_s}{\Delta E}\right) \tag{5-3-37}$$

This is a Fourier series describing ϵ in terms of E_s. For $\sigma = 0$, this becomes the series representation of a sawtooth waveform with period ΔE and peak amplitude $\pm \Delta E/2$. For $\sigma \gg \Delta E$, the only term of significance is $m = 1$. Thus,

$$\epsilon \approx -\frac{\Delta E}{\pi} e^{-2(\pi\sigma/\Delta E)^2} \sin\left(2\pi \frac{E_s}{\Delta E}\right) \tag{5-3-38}$$

$$|\epsilon|_{\max} = \frac{\Delta E}{\pi} e^{-2(\pi\sigma/\Delta E)^2} \tag{5-3-39}$$

For example, suppose $|\epsilon|_{\max} = 0.01(\Delta E/2)$. Then $\sigma = 0.4587\Delta E$. Thus two samples within the gaussian noise envelope will give a mean-value

error reduction of 100. The variance can be written as

$$\overline{E^2} - (\bar{E})^2 = \frac{1}{n}\left[\frac{1}{12}(\Delta E)^2 + \sigma^2\right] + \delta^2 \approx \frac{1}{n}\left[\frac{1}{12}(\Delta E)^2 + \sigma^2\right] = \Sigma^2 \tag{5-3-40}$$

$$\delta^2 \approx -\frac{1}{n} e^{-2(\sigma\pi/\Delta E)^2}\left[\left(\frac{\Delta E}{\pi}\right)^2 \cos 2\pi \frac{E_s}{\Delta E} + 4\sigma^2 \cos 2\pi \frac{E_s}{\Delta E} - \frac{\Delta E\, Es}{\pi} \sin 2\pi \frac{E_s}{\Delta E}\right] \tag{5-3-41}$$

The expression for δ^2 assumes $m = 1$. Note that δ^2 is very small because of the term $e^{-2(\sigma\pi/\Delta E)^2}$ and so will be neglected. Thus the mean and second moment of the final voltage after sampling have been determined. For a specified number of averages, n, and a specified confidence level A, how much noise σ should be added prior to sampling to reduce the digitizing error to a minimum, and what is this minimum value? The optimum σ can be obtained by differentiating A with respect to σ and equating to zero. The quantity $|\epsilon|_{\max}$ is used in place of ϵ. The result of differentiation is

$$\coth\left(\frac{\epsilon E_0}{\Sigma^2}\right) = \frac{\epsilon}{E_0} - \frac{\Sigma}{E_0}\left(\frac{d\epsilon/d\sigma^2}{d\Sigma/d\sigma^2}\right) = \frac{\Delta E}{\pi E_0}\left[1 + (2\pi)^2\left(\frac{1}{12} + \frac{\sigma^2}{\Delta E^2}\right)\right] e^{-(\pi\sigma/\Delta E)^2} \tag{5-3-42}$$

Define

$$x = \frac{\epsilon E_0}{\Sigma^2} \quad \text{and} \quad k = \frac{E_0}{(2)^{1/2}\Sigma}$$

Then

$$x \tanh x = \frac{2k^2}{1 + (2\pi)^2[\frac{1}{12} + (\sigma^2/\Delta E^2)]} \tag{5-3-43}$$

where

$$x = \frac{nE_0}{\pi\,\Delta E} \frac{1}{[\frac{1}{12} + (\sigma^2/\Delta E^2)]} e^{-2(n\sigma/\Delta E)^2}$$

and

$$k = \frac{E_0}{\Delta E}\left(\frac{n}{2}\right)^{1/2} \frac{1}{[\frac{1}{12} + (\sigma^2/\Delta E^2)]^{1/2}}$$

The confidence level A can be expressed as

$$A = \tfrac{1}{2}\left[\operatorname{erf}\left(k + \frac{x}{2k}\right) + \operatorname{erf}\left(k - \frac{x}{2k}\right)\right] \tag{5-3-44}$$

Normally $|x/2k| \ll |k|$ and so A may be approximated as follows:

$$A \approx \operatorname{erf}(k) - \frac{x^2}{2k(\pi)^{1/2}} e^{-k^2} \tag{5-3-45}$$

Assume that A is given, and evaluate an approximation to k by using $A \approx \operatorname{erf}(k)$. For a given value of $\sigma/\Delta E$ then, x is calculated by Eq. (5-3-43). This is a transcendental equation, but it can be readily solved on a slide rule with the D and Th scales. These values of k and x are inserted in Eq. (5-3-45) to obtain a closer value of k. Normally one iteration is sufficient. By using the final values of k and x, it is possible to calculate n and E_0:

$$n = \frac{1}{2}\left(\frac{\pi x}{k}\right)^2 \left(\frac{1}{12} + \frac{\sigma^2}{\Delta E^2}\right) e^{4(\pi\sigma/\Delta E)^2} \tag{5-3-46}$$

$$\frac{E_0}{\Delta E} = \frac{2k^2}{\pi x} e^{-2(n\sigma/\Delta E)^2} \tag{5-3-47}$$

Because of the approximation $m = 1$, this method begins to lose accuracy for $\sigma/\Delta E < 0.4$ ($n < 100$ or $E_0/\Delta E > 0.1$). Curves of n and $E_0/\Delta E$ versus $\sigma/\Delta E$ are shown in Fig. 5-14. For example, if $\sigma/\Delta E = 0.57$ and $n = 10^5$, then $E_0/\Delta E = 0.005$. This represents a reduction in analog-to-digital converter error by a factor of 100. Note that only a very small amount of additive noise is needed prior to sampling.

Spectral Analysis. Perhaps the most important tool in signal analysis is the Fourier transform. This is primarily true because the Fourier transformation occurs so often in nature. Any time an observation involves a linear summation of phasors of arbitrary amplitude, the Fourier transform is included. Linear differential equations have solutions using the transform of the driving function. Optical, electromagnetic, and acoustical far-field distributions are transforms of the source function. A time waveform and its frequency spectrum are Fourier transform pairs. The primary characteristic of the transform that makes it so useful is related to the inverse nature of the transform variables. Fine structure in a time waveform, for instance, becomes gross structure in the transform domain. This implies a presentation of the data in an entirely different form, and allows new observations which might otherwise be hidden.

The nature of the discrete Fourier transform has already been discussed, and the concept of a time window and a line shape have been introduced.

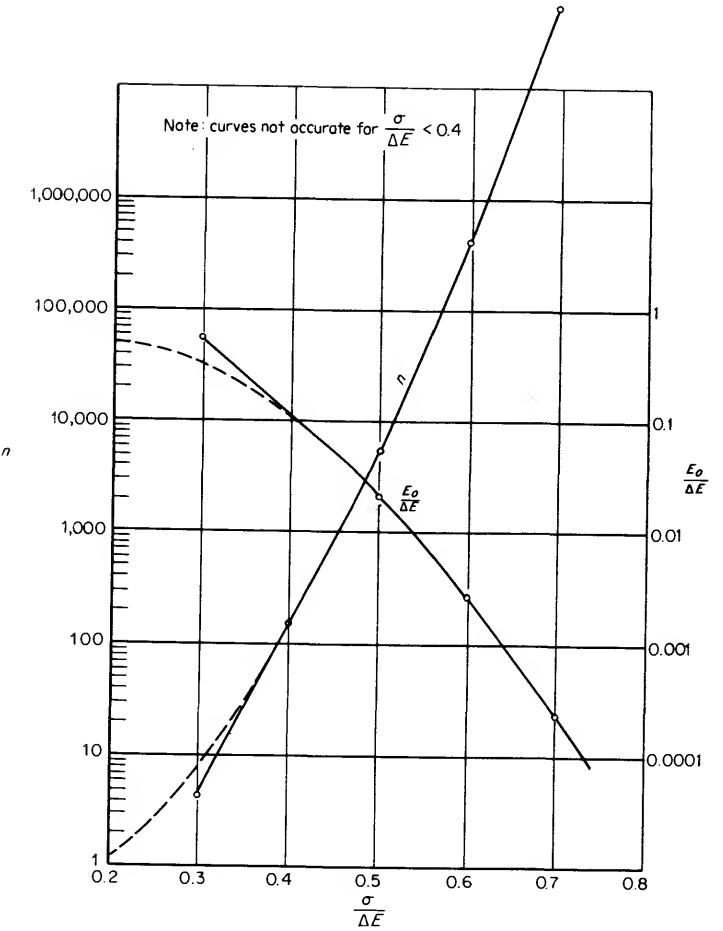


FIG 5-14 Optimum values of noise, $\sigma/\Delta E$, and resultant error, $E_0/\Delta E$ for a given number of averages, n, for a confidence level A of 0.99.

These are fundamental to the spectral analysis process and will be investigated somewhat more thoroughly. The time window and its associated line shape are Fourier transform pairs. A finite time record of data is usually obtained by multiplying an infinite record by a rectangular time window. However, the resulting sin $\pi sT/\pi s$ line shape has serious disadvantages. The side lobes only drop 6 dB/octave, and hence adjacent frequency channels interfere with one another (see Fig. 5-15*a*).

In order to reduce this interference (often called *leakage*) between channels, it is necessary to generate a line shape with smaller side lobes [8]. In the time domain, this corresponds to a window whose transitions are less abrupt than the rectangular ones. The data is weighted or attenuated near both ends of the time interval. This will broaden the

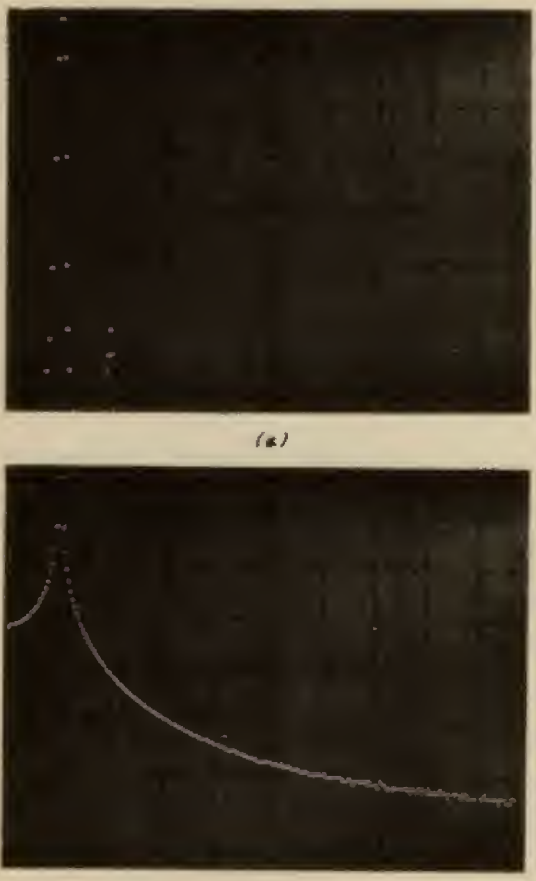


FIG 5-15 Results of Hanning on Fourier analysis: (a) Positive frequency spectrum of sum of sine wave and second harmonic. Sine wave has 25.5 cycles in time window. Second harmonic is 60 dB below fundamental. Note hyperbolic line shape and "leakage" between harmonics. (b) Same as above except the Hanning operation was performed twice. Note the separation between the two harmonics. The vertical scale is 10 dB/cm.

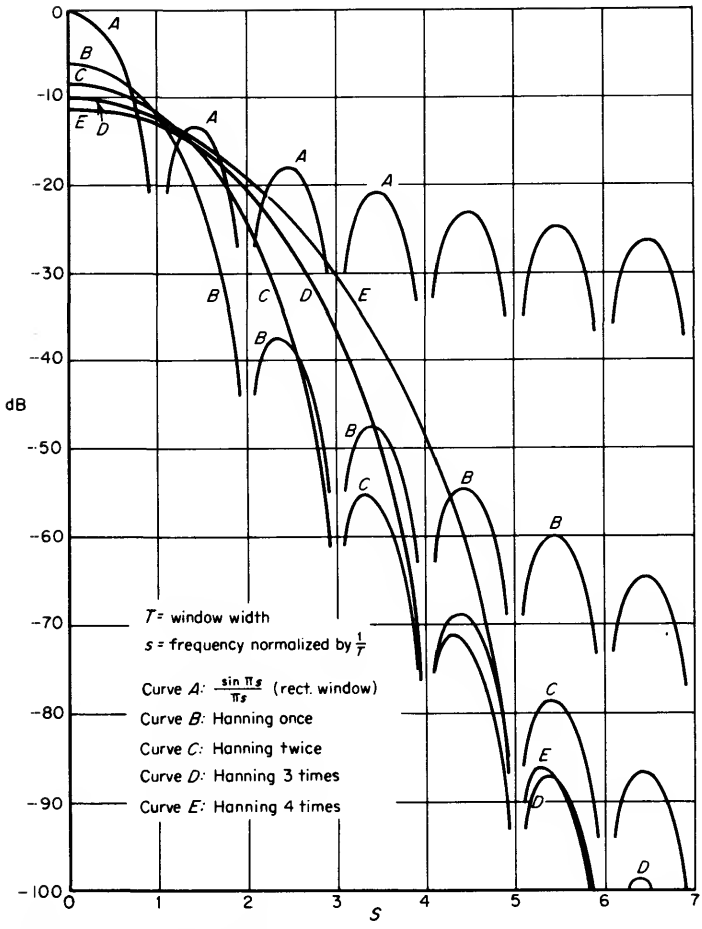


FIG 5-16 Hanning line shapes.

main lobe of the line shape and thereby reduce resolution in the frequency domain, but this is the price for better separation between adjacent harmonics. There are many window functions that can be used, and numerous shapes are described in the literature on this subject. Each one has advantages and disadvantages depending upon the desired result. In this chapter, only two will be discussed in detail: (1) the Hanning window, and (2) the Chebyshev window. Hanning is useful because it is so easy to implement, and the Chebyshev window is of interest because it gives the narrowest main lobe for a given side-lobe amplitude.

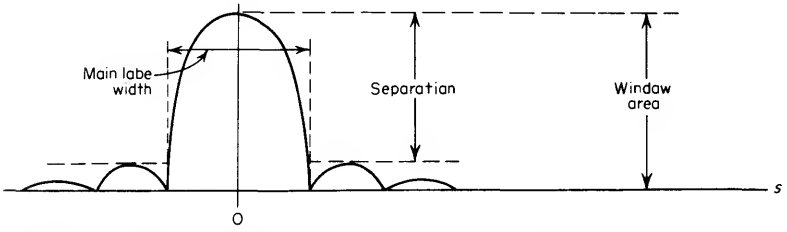


FIG 5-17 Definitions applying to a line shape in the frequency domain.

The Hanning window is $g(t) = \frac{1}{2}[1 - (\cos 2\pi t/T)]$ for $0 \leq t \leq T$. The corresponding sampled line shape is

$$G(s) = \tfrac{1}{2}\delta(s) - \tfrac{1}{4}\delta(s - 1) - \tfrac{1}{4}\delta(s + 1)$$

Thus the Hanning window can be easily implemented by convolving the frequency function with a triplet of delta functions with binary amplitudes. Each Hanning operation attenuates the side lobes by at least 12dB/octave (see Fig. 5-16). The height of the window is unity, but the area has been reduced by 2. This means that discrete spectral lines will retain their area, but their amplitude will be reduced. Main-lobe width is defined as the frequency interval between points on the main lobe that have the same amplitude as the absolute magnitude of the largest side lobe. Separation is defined as the ratio of maximum main-lobe magnitude to maximum side-lobe magnitude. Figure 5-17 shows these relationships.

Table 5-2 shows the main-lobe width, separation, and window area for various Hanning operations. Graphs are included in Fig. 5-16. Recall that $\Delta s = 1/T$ is the spacing between adjacent points on the frequency axis. The relative window area indicates the reduction in height of discrete spectral lines, caused by the line shape's smearing. Note that the separation improves very rapidly in comparison with the degradation of main-lobe width and window area.

TABLE 5-2

Function	Separation, dB	Main-lobe width	Relative window area, dB
Rectangular window	13.26	$1.626/T$	0
Hanning once	31.47	$3.743/T$	−6.02
Hanning twice	46.74	$5.782/T$	−8.52
Hanning 3 times	60.95	$7.804/T$	−10.10
Hanning 4 times	74.61	$9.818/T$	−11.26

The Chebyshev window (Fig. 5-18) is more difficult to implement than Hanning's, but it does a better job. The name is used because of the equal ripple side lobes of the line shape. The line shape is given by

$$G(s) = \frac{\cos\left[(\pi s T)^2 - a^2\right]^{1/2}}{aI_1(a)} \qquad a = \cosh^{-1}\frac{1}{r} \tag{5-3-48}$$

where r is the ripple factor, or separation between the main lobe and the side lobes, and $I_1(a)$ is the modified Bessel function of unit order. The

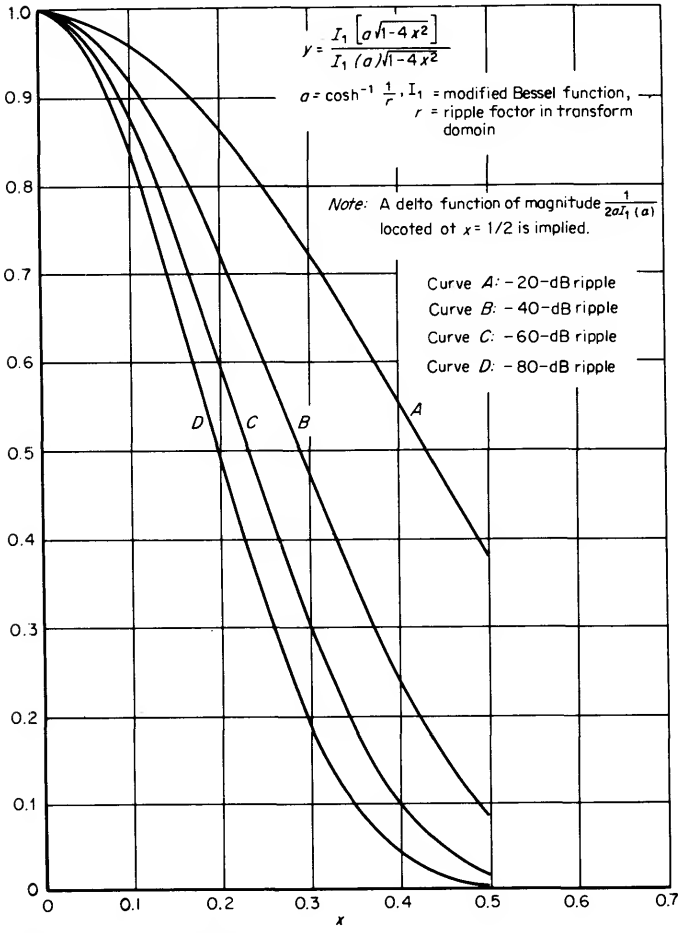


FIG 5-18 Chebyshev window.

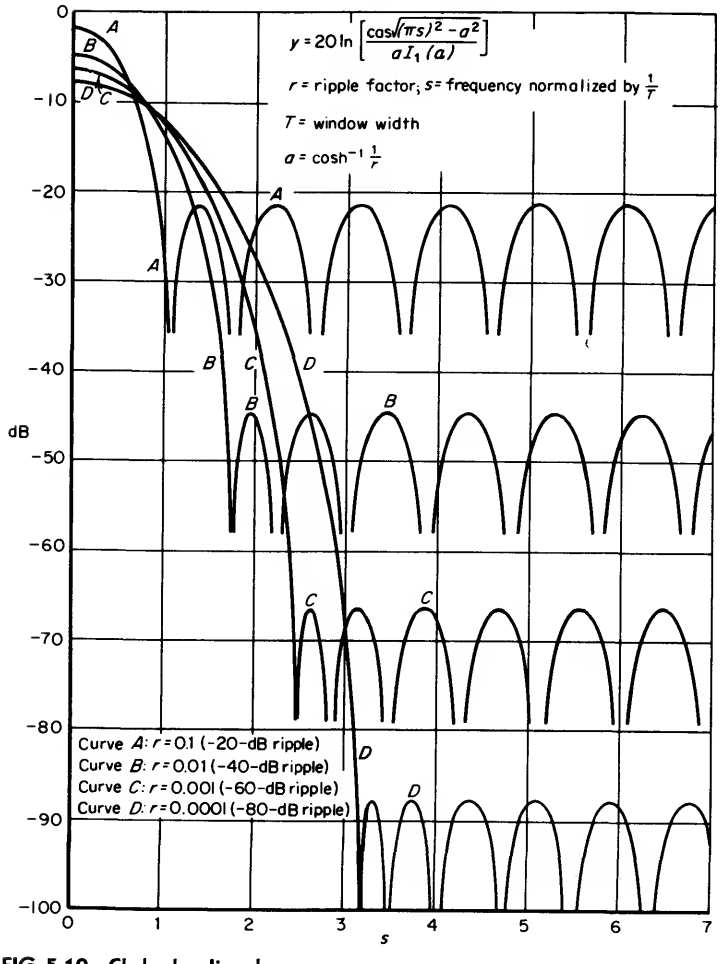


FIG 5-19 Chebyshev line shapes.

corresponding time window is

$$g(t) = \frac{1}{2aI_1(a)} \left[\delta\left(\frac{t}{T} + \frac{1}{2}\right) + \delta\left(\frac{t}{T} - \frac{1}{2}\right) + 2a \frac{I_1(a\sqrt{1 - (2t/T)^2})}{\sqrt{1 - (2t/T)^2}} \right] \tag{5-3-49}$$

Graphs of this line shape (Fig. 5-19) for various ripple factors are included. Table 5-3 summarizes the characteristics of this line shape.

TABLE 5-3 Characteristics of Chebyshev Window

Ripple factor, dB	Separation, dB	Main-lobe width	Relative window area, dB
−20	20	$1.906/T$	−1.40
−40	40	$3.373/T$	−4.59
−60	60	$4.839/T$	−6.38
−80	80	$6.305/T$	−7.64
−100	100	$7.771/T$	−8.62

The window height is unity and therefore the line-shape area is unity. The line-shape height and the window area are $(\cosh a)/[aI_1(a)]$. The peak ripple is $1/[aI_1(a)]$. The window function includes delta functions of amplitude $1/[2aI_1(a)]$ at each end of the window interval. The window amplitude immediately adjacent to these end points is $a/[2I_1(a)]$.

A graph of separation versus main-lobe width (Fig. 5-20) is included for comparison of various window shapes. Note that the Chebyshev line shape gives considerably higher resolution for a given separation. The Parzen window given by $g(t) = 1 - 2|t|$ is also shown on this plot. This window has a relative area of −6.02 dB and a line shape given by

$$G(s) = 2\left[\frac{\sin(\pi s/2)}{\pi s}\right]^2$$

The separation is 26.52 dB and the main-lobe width is $3.251/T$. This is twice the separation and half the resolution obtained with a rectangular window.

The rectangular window is fine for signals that have an integer number of periods within the window interval T, or for single transients which decay within the T interval. A weighting function is only needed when the waveform value at the beginning of the window is different from the value at the end. This is particularly important for random data and for coherent data of arbitrary frequency. All spectral lines are convolved (smeared) with the line shape, and thus the amplitudes of narrow spectral lines are reduced. The magnitudes of the various spectral components are changed by this smearing operation, but the areas (and hence voltage or power) under each line are preserved. It is difficult to accept the loss in frequency resolution that these line shapes produce, but the only alternative is very poor amplitude accuracy because of poor separation between adjacent frequencies (see Fig. 5-16).

One word of caution is in order concerning the cumulative effect of line-shape side lobes. If the original data contain several discrete frequencies, then the line shape will be convolved about each spectral

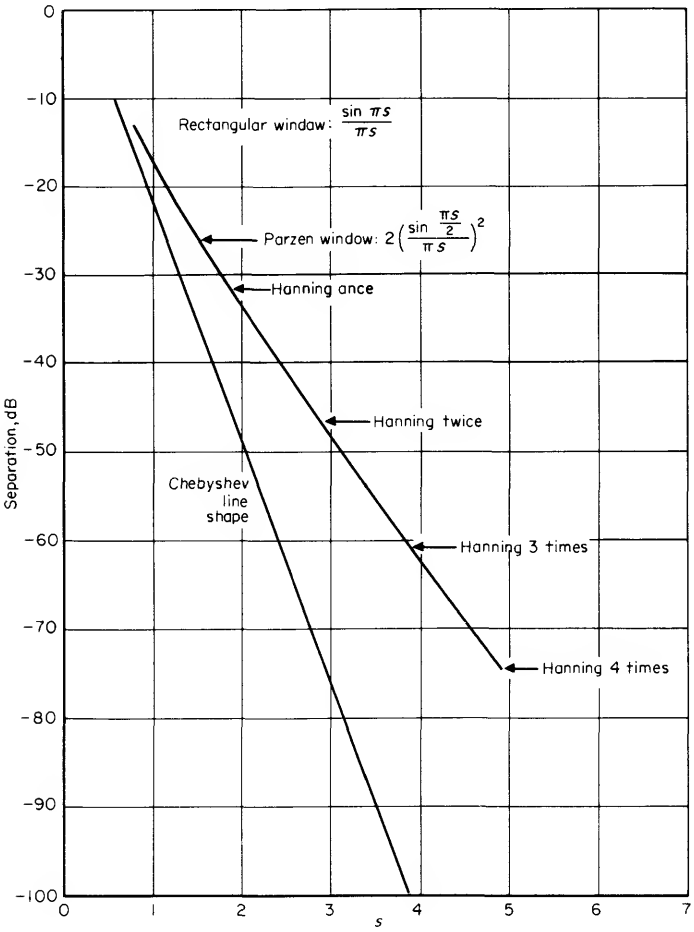


FIG 5-20 Separation vs. lobe width. T = **window width,** s = **frequency, normalized to** $1/T$.

line. It is possible for the line-shape ripple to add in phase in some regions of the frequency domain. Therefore the resultant ripple may greatly exceed the amount calculated for a single spectral line. The Chebyshev line shape is particularly bad in this respect since the line-shape ripple extends uniformly across the frequency scale.

Power Spectral Analysis. The power spectrum of a signal has some properties that are often very useful. Since phase information is lost, there is no need for sampling in synchronism with any signal. The power spectrum is meaningful for random data as well as for coherent

signals. Thus it is useful for searching unknown data to determine the nature of any information that may be present. Since translation of the time origin has no effect on the power spectrum, the spectrum is often used as a "signature" to represent a particular process.

The power spectrum of a coherent signal is simply the square of the magnitude of the Fourier transform of that signal. However, if the signal is random, it is necessary to accumulate an ensemble average of the power spectra from many data records. If nothing is known about the coherent nature of the signal, then a single power spectrum is essentially meaningless. It is difficult to distinguish a coherent spectral line from a random line with only a single record.

It is generally impractical to perform a true ensemble average since this requires a large number of records from identical sources. Instead, it is common to assume an ergodic process and to use many records taken at different times from a single source. Unfortunately there are many processes which are not stationary (and therefore not ergodic). A typical example is so-called "$1/f$ noise" from semiconductors. These nonstationary processes may prevent the power spectrum from converging to a stable value at each frequency as various time records are averaged.

For ergodic processes the power spectrum does converge and the "stability" of the result after a finite number of averages may be calculated. As indicated previously, the process of averaging a random variable leaves the mean unchanged but reduces the variance by the number of points that are averaged. Thus, the first step is to calculate the moments and probability density of the power spectrum at each frequency point in the absence of signal averaging. In order to make the mathematics more explicit, assume that both the real and imaginary parts of the Fourier transform of the signal are independent gaussian random variables. This will be the case if the original time-domain variable is gaussian. Define the following:

$$z = x^2 + y^2 \tag{5-3-50}$$

where x and y are independent random variables representing the real and imaginary parts of the signal spectrum, and z is the random variable representing the power spectrum at each frequency.

$$f_{xy}(x,y) = \frac{1}{2\pi\sigma^2} e^{-(1/2\sigma^2)\,[(x-\bar{x})^2+(y-\bar{y})^2]} \tag{5-3-51}$$

This is the joint probability density of x and y. Both variables have the same variance σ^2, and their respective mean values are $\bar{x}$ and $\bar{y}$. These mean values represent the coherent part of the Fourier transform at each frequency. The probability density of z and the various moments will be calculated.

$$f_z(z)\,dz = f_{xy}(x,y)\,dA \tag{5-3-52}$$

where $f_z(z)$ is the probability density of z, and dA is the area element in the xy plane corresponding to the interval dz. In polar coordinates: $x = r \cos \theta$, $y = r \sin \theta$, $z = r^2$, $dz = 2r\,dr$, and $dA = r\,dr\,d\theta = \frac{1}{2}dz\,d\theta$. Thus,

$$f_z(z) = \frac{1}{4\pi\sigma^2} e^{-(1/2\sigma^2)(z+\bar{x}^2+\bar{y}^2)} \int_{-\pi}^{\pi} e^{(\sqrt{z}/\sigma^2)(\bar{x}\cos\theta+\bar{y}\sin\theta)}\,d\theta \tag{5-3-53}$$

Some algebraic manipulation allows the integral to be written in a tabulated form. See "Handbook of Mathematical Functions," National Bureau of Standards, page 376, Eq. (9.6.16).

$$I_0(\omega) = \frac{1}{\pi}\int_0^{\pi} e^{\pm\omega\cos\phi}\,d\phi \tag{5-3-54}$$

Thus $f_z(z)$ can be written in closed form as

$$f_z(z) = \frac{1}{2\sigma^2} I_0\left(\frac{z^{1/2}}{\sigma^2}(\bar{x}^2+\bar{y}^2)^{1/2}\right) e^{-(1/2\sigma^2)(z+\bar{x}^2+\bar{y}^2)} \qquad \text{for } z \geq 0 \tag{5-3-55}$$

It is also possible to derive the characteristic function $F_z(s)$,

$$F_z(s) = \int_0^{\infty} f_z(z) e^{-i2\pi sz}\,dz \tag{5-3-56}$$

From the same reference as above, page 486, Eq. (11.4.29), the following integral can be derived:

$$\int_0^{\infty} \omega e^{-a^2\omega^2} I_0(b\omega)\,d\omega = \frac{1}{2a^2} e^{b^2/4a^2} \tag{5-3-57}$$

Substituting $z = \omega^2$, $dz = 2\omega\,d\omega$, and expressions for a and b,

$$F_z(s) = \frac{1}{1 + i4\pi\sigma^2 s} e^{-(\bar{x}^2+\bar{y}^2)[i2\pi s/(1+i4\pi\sigma^2 s)]} \tag{5-3-58}$$

By successive differentiation it is possible to derive the various moments of $f_z(z)$ that are of interest, as follows:

$$\text{Area under } f(z) = F_z(0) = 1 \tag{5-3-59}$$

$$\text{Mean value } \bar{z} = \frac{F_z'(0)}{-i2\pi} = \bar{x}^2 + \bar{y}^2 + 2\sigma^2 \tag{5-3-60}$$

$$\text{Second moment } \overline{z^2} = \frac{F_z''(0)}{-4\pi^2} = (\bar{x}^2 + \bar{y}^2)^2 + 8\sigma^2(\bar{x}^2 + \bar{y}^2 + \sigma^2) \tag{5-3-61}$$

$$\text{Variance } \overline{z^2} - (\bar{z})^2 = 4\sigma^2(\bar{x}^2 + \bar{y}^2 + \sigma^2) \tag{5-3-62}$$

The coherent signal power is $\bar{x}^2 + \bar{y}^2$ and the noise power is $2\sigma^2$ (since σ^2

occurs in both real and imaginary parts). Thus the mean of z is the total power. The variance describes the statistical variation about this mean value.

For the special case of $\bar{x}^2 + \bar{y}^2 = 0$,

$$f_z(z) = \frac{1}{2\sigma^2} e^{-z/2\sigma^2} \qquad \text{for } z \geq 0 \tag{5-3-63}$$

$$F_z(s) = \frac{1}{1 + i4\pi\sigma^2 s} \tag{5-3-64}$$

$$\bar{z} = 2\sigma^2 \tag{5-3-65}$$

$$\overline{z^2} = 8\sigma^4 \tag{5-3-66}$$

$$\overline{z^2} - \bar{z}^2 = 4\sigma^4 \tag{5-3-67}$$

Note that the standard deviation (square root of the variance) is equal to the mean. This is a chi-squared distribution with two degrees of freedom.

The average of n records will normally have a gaussian probability density (central-limit theorem) and a variance equal to the above value divided by n. The mean is unchanged.

$$f_{z_n}(z_n) = \frac{1}{\sigma_n (2\pi)^{1/2}} e^{-1/2\,[(z_n - \bar{z})/\sigma_n]^2} \tag{5-3-68}$$

where

$$\sigma_n = \frac{2\sigma}{n^{1/2}} (\bar{x}^2 + \bar{y}^2 + \sigma^2)^{1/2}$$

and

$$\bar{z} = \bar{x}^2 + \bar{y}^2 + 2\sigma^2$$

$$F_{z_n}(s) = e^{-2(\sigma_n \pi s)^2} e^{-i2\pi \bar{z} s} \tag{5-3-69}$$

$$\overline{z_n} = \bar{z} = \bar{x}^2 + \bar{y}^2 + 2\sigma^2 \tag{5-3-70}$$

$$\overline{z_n^2} - \bar{z}_n^2 = \sigma_n^2 = \frac{4\sigma^2}{n} (\bar{x}^2 + \bar{y}^2 + \sigma^2) \tag{5-3-71}$$

$$\frac{\sigma_n}{\bar{z}_n} n^{1/2} = \frac{2r(1 + r^2)^{1/2}}{1 + 2r^2} \tag{5-3-72}$$

where

$$r = \frac{\sigma}{(\bar{x}^2 + \bar{y}^2)^{1/2}}$$

The last equation expresses the ratio of standard deviation to mean value of each point in a power density spectrum as a function of the noise-to-signal ratio r of the original frequency spectrum and the number of records, n, that are averaged. A graph of this function is included in Fig. 5-21. Note that for large amounts of noise, the standard deviation

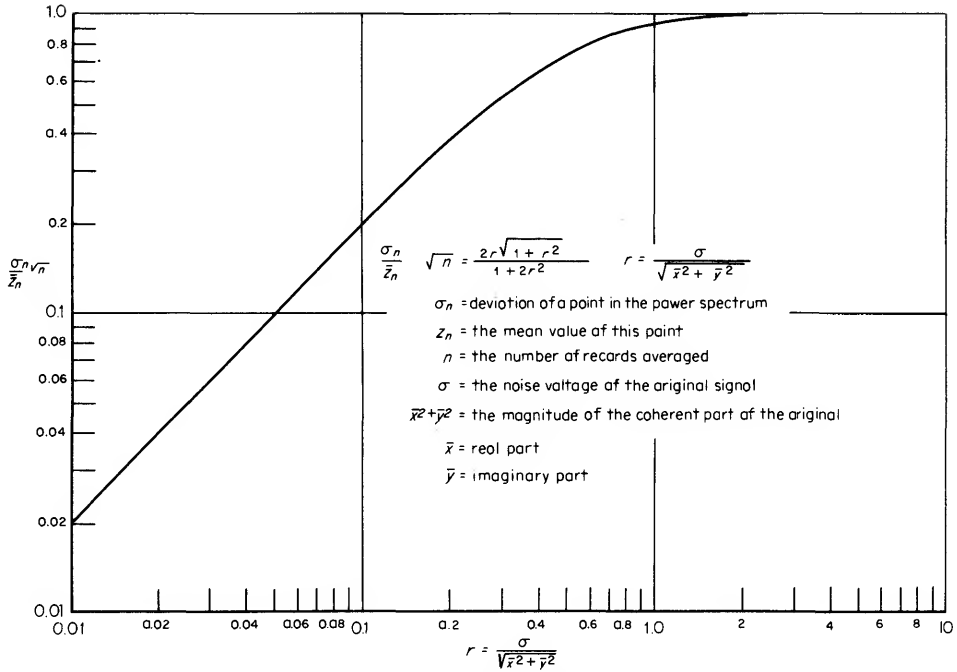


FIG 5-21 Relative uncertainty in power spectrum vs. noise-to-coherent signal ratio.

σ_n of the power at each frequency is simply the mean power value divided by $\sqrt{n}$.

It is of considerable importance to determine the possible resolution of a coherent spectral line buried in noise by using the power averaging technique. Consider a random variable $\omega = z_n - z_{no}$, where z_{no} is the random variable for noise alone ($\bar{x}^2 + \bar{y}^2 = 0$). Then $\bar{\omega} = \bar{z}_n - \bar{z}_{no} = \bar{x}^2 + \bar{y}^2$ and $\sigma_\omega^2 = \sigma_n^2 + \sigma_{no}^2 = (4\sigma^2/n)(\bar{x}^2 + \bar{y}^2 + 2\sigma^2)$. Note that $\bar{\omega}$ represents the strength of the coherent spectral line, while σ_ω^2 is the variance about this mean value. Hence, if ω is assumed to be a gaussian random variable, confidence limits on the determination of $\bar{x}^2 + \bar{y}^2$ can be readily established. For example, the probability is 0.99 that a coherent spectral line will exceed a neighboring noise line if $\bar{\omega} = k\sigma_\omega$ for $k = 2.327$. This implies that $n = 4k^2r^2(2r^2 + 1)$. For unity noise-to-signal ratio $r = 1$, and $n = 12k^2 \approx 65$. Note that for large values of n, the term r is proportional to the fourth root of n.

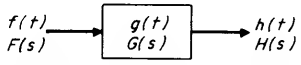


FIG 5-22 Generalized filter situation.

It may sometimes be more convenient to average the amplitude of a frequency spectrum instead of the power. This solution is discussed in the appendix to this chapter. In general, amplitude averaging is not quite as efficient and is more difficult to interpret in comparison with power averaging.

Filtering and Convolution. The concept of filtering in the frequency domain (or convolution in the time domain) is a very fundamental and useful one. Consider an arbitrary system with an input and an output signal, as shown in Fig. 5-22. The input signal $F(s)$ is multiplied by the system transfer function $G(s)$ to obtain the output signal $H(s)$. There are three distinct measurement situations in which any two of these signals are known and the third one is desired. These will be considered one at a time.

1. $F(s)$ and $G(s)$ known. Find $H(s)$.

This is a conventional filtering operation where $H(s) = F(s)G(s)$. In the time domain, this corresponds to convolution between $f(t)$ and $g(t)$. The term $H(s)$ is often called the *cross spectrum* between $F(s)$ and $G(s)$. Only those spectral lines that are common to $F(s)$ and $G(s)$ will occur in the output. Here $G(s)$ can be any conventional analog filter function, although many other filters that are physically unrealizable can be readily implemented by the digital approach. For example, it is easy to attenuate signal amplitude without introducing any phase shift. Rectangular filters become very simple, and can be made very narrow to select only one frequency.

2. $F(s)$ and $H(s)$ known. Find $G(s)$.

This is simply the measurement of the transfer function of a device or system. The equation $G(s) = H(s)/F(s)$ can be implemented digitally with considerable accuracy. The division operation can introduce errors if $F(s)$ is not accurately known, so it is desirable to keep $F(s)$ as large as possible at all frequencies. This operation is called *inverse filtering* or *deconvolution*, and it will be discussed more fully below.

3. $G(s)$ and $H(s)$ known. Find $F(s)$.

This is the most general deconvolution operation. The equation $F(s) = H(s)/G(s)$ is analogous to the above situation, but the errors introduced by lack of control of $G(s)$ may now be very significant. This operation is used to attempt the reconstruction of an original signal $f(t)$ after it has passed through a known filter $G(s)$. The technique can be used to improve resolution in some measurements, or it can be used to equalize a prescribed transmission path to allow recovery of the original waveform.

The problem is simply that $G(s)$ generally becomes very small or zero for some values of s. For the regions where $G(s) = 0$, all information about $F(s)$ is completely missing and can never be recovered. Furthermore, even if $G(s) \neq 0$ but small, there is always "noise" on both $G(s)$ and $H(s)$ because of either random fluctuations in the data or finite resolution of numbers within the measuring instrument. When $G(s)$ falls below this noise level, large uncertainties are introduced into the determination of $F(s)$. In some situations $G(s)$ may have local zeros which will introduce local regions of uncertainty in $F(s)$. Often it may safely be assumed that $F(s)$ is continuous across these uncertain regions. Although this technique must be used with care, it does allow the use of data beyond the first zero of $G(s)$.

It is necessary to differentiate between the expected and the actual values of these frequency functions. For *each* value of s, all three functions F, G, and H will be random variables with probability densities $p_F(F)$, $p_G(G)$, and $p_H(H)$. The means of these distributions will be denoted by $\bar{F}$, $\bar{G}$, and $\bar{H}$.

The basic problem can now be stated: What is the "best" estimate of $F(s)$ obtainable from the *actual* quotient $H(s)/G(s)$?

The question of what constitutes a best fit to a noisy function is debatable and depends to a considerable extent on the application. In this discussion, a linear mean-square estimate will be assumed, based on the Wiener-Kolmogoroff theory. See Ref. 9, pages 400 to 405, for further treatment. The technique is to pass the random variable

$$F(s) = H(s)/G(s)$$

through a "matched" filter, denoted by $\phi(s)$, whose shape is given by

$$\phi(s) = \frac{\bar{F}(s)\bar{F}^*(s)}{\overline{|F(s)|^2}} \tag{5-3-73}$$

where $\bar{F}$ is the mean and $F^*(s)$ is the complex conjugate of the variable $F(s)$. The bars denote ensemble averaging. Thus the best estimate of $F(s)$ in the context of this discussion is given by

$$\hat{F}(s) = \frac{H(s)}{G(s)}\phi(s) = F(s)\phi(s) \tag{5-3-74}$$

The construction of a matched filter requires some knowledge about the function $F(s)$. Essentially it is necessary to know either the coherent part or the random part of $F(s)$. Thus the construction of a matched filter is something of a bootstrap operation. The output of this filter tends to reinforce the assumptions used to create the matched-filter function in the first place. Therefore, the results must be interpreted with care. Assume that $F(s)$ consists of a coherent part $\bar{F}(s)$ and a noise part $N(s)$,

$$F(s) = \bar{F}(s) + N(s) \tag{5-3-75}$$

where $\overline{N(s)} = 0$. Then the total signal power is

$$\overline{|F(s)|^2} = |\bar{F}(s)|^2 + \overline{N^2(s)} \tag{5-3-76}$$

and the coherent signal power is

$$\bar{F}(s)\bar{F}^*(s) = |\bar{F}(s)|^2 \tag{5-3-77}$$

$$\phi(s) = \frac{|\bar{F}(s)|^2}{|\bar{F}(s)|^2 + \overline{N^2(s)}} = 1 - \frac{\overline{N^2(s)}}{|\bar{F}(s)|^2 + \overline{N^2(s)}} \tag{5-3-78}$$

where $\overline{N^2(s)}$ is the noise power. These equations all assume that the noise is random and completely uncorrelated with the coherent signal $\bar{F}(s)$.

The calculation of the noise power $\overline{N^2(s)}$ will be considered next. This is simply the variance of the random variable F, but $F = H/G$. Thus it is necessary to calculate the variance of the quotient of two random variables. For another discussion, see Ref. 9, pages 196 to 197. The basic formula is

$$p_F(F) = \int_{-\infty}^{\infty} |\lambda| p_H(\lambda F) p_G(\lambda)\, d\lambda \tag{5-3-79}$$

where G and H are assumed to be independent random variables.

Although this integral could be evaluated if necessary, only the first and second moments are needed here.

$$\bar{F} = \int_{-\infty}^{\infty} F p_F(F)\, dF = \int_{-\infty}^{\infty} \frac{|\lambda|}{\lambda^2} p_G(\lambda) \int_{-\infty}^{\infty} \lambda F p_H(\lambda F)\, d(\lambda F)\, d\lambda \quad (5\text{-}3\text{-}80)$$

However,

$$\bar{H} = \int_{-\infty}^{\infty} \lambda F p_H(\lambda F)\, d(\lambda F)$$

Thus,

$$\bar{F} = \bar{H} \int_{-\infty}^{\infty} \frac{|\lambda|}{\lambda^2} p_G(\lambda)\, d\lambda \quad (5\text{-}3\text{-}81)$$

In a similar manner,

$$\overline{F^2} = \overline{H^2} \int_{-\infty}^{\infty} \frac{|\lambda|}{\lambda^3} p_G(\lambda)\, d\lambda \quad (5\text{-}3\text{-}82)$$

Note that only the first and second moments of H are needed, although for G the entire density function is required.

Consider a practical example where the uncertainty in G is caused by digital roundoff errors. Assume a rectangular probability density for G as follows:

$$p_G(G) = \frac{1}{a} \sqcap \left(\frac{G - \bar{G}}{a} \right) \quad (5\text{-}3\text{-}83)$$

The distribution is of width a, centered at $\bar{G}$. Further, assume that $\bar{G} > a/2$. Then

$$\bar{F} = \bar{H} \int_{\bar{G}-a/2}^{\bar{G}+a/2} \frac{d\lambda}{a\lambda} = \frac{2\bar{H}}{a} \tanh^{-1} \frac{a}{2\bar{G}} \quad (5\text{-}3\text{-}84)$$

$$\overline{F^2} = \overline{H^2} \int_{\bar{G}-a/2}^{\bar{G}+a/2} \frac{d\lambda}{a\lambda^2} = \frac{\overline{H^2}}{\bar{G}^2 - \frac{1}{4}a^2} \quad (5\text{-}3\text{-}85)$$

The noise power is given by the variance of F, that is, $\overline{N^2(s)} = \overline{F^2} - \bar{F}^2$. Thus the matched filter can be written as

$$\phi(s) = \frac{\bar{F}^2}{\overline{F^2}} = \frac{4\bar{H}^2}{a^2 \overline{H^2}} (\bar{G}^2 - \tfrac{1}{4}a^2) \left(\tanh^{-1} \frac{a}{2\bar{G}} \right)^2 \quad (5\text{-}3\text{-}86)$$

Assume that the original input function $f(t)$ was a very narrow pulse or delta function. Then $\bar{F}(s) =$ constant. Also assume that $\overline{H^2} = \bar{H}^2$, so

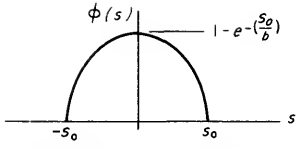


FIG 5-23 The matched filter function $\phi(s)$.

that H is noise free. Then,

$$\phi(s) = .1 - \left(\frac{a}{2\bar{G}}\right)^2 \qquad \text{for } \bar{G} > \frac{a}{2} \tag{5-3-87}$$

Now let $g(t)$ be a gaussian smearing function so that $\bar{G}(s)$ is given by

$$\bar{G}(s) = ke^{-\frac{1}{2}(s/b)^2} \tag{5-3-88}$$

$$\phi(s) = 1 - \left(\frac{a}{2k}\right)^2 e^{(s/b)^2} \qquad \text{for } k \gg \frac{a}{2} \tag{5-3-89}$$

Define s_0 as the zero of $\Phi(s)$,

$$s_0 = b\left(2 \ln \frac{2k}{a}\right)^{\frac{1}{2}} \tag{5-3-90}$$

$$\frac{2k}{a} = e^{\frac{1}{2}(s_0/b)^2} \tag{5-3-91}$$

$$\phi(s) = 1 - e^{(s^2-s_0^2)/b^2} \text{ for } |s| \leq s_0 \tag{5-3-92}$$

The graph of this function is shown in Fig. 5-23.

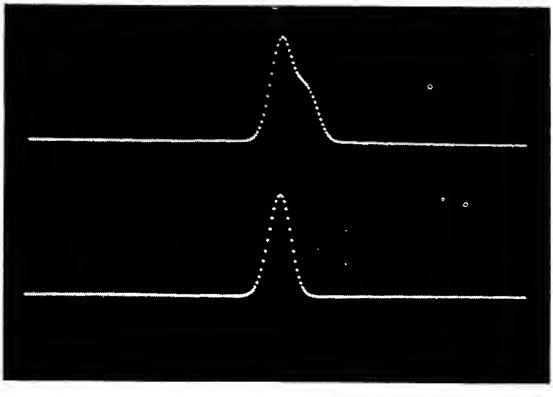


FIG 5-24 (*a*) Output function: convolution of Gaussian "smearing" function with pair of delta functions. (*b*) Gaussian "smearing" function.

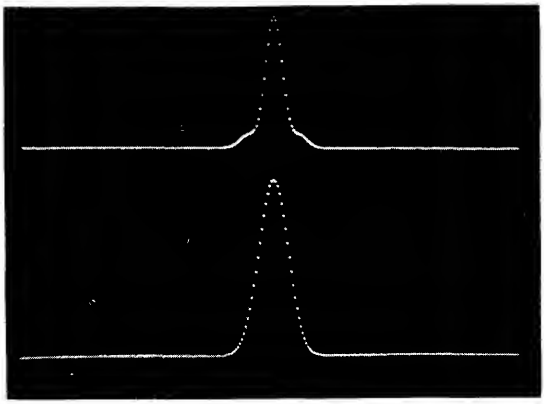


FIG 5-25 (*a*) Fourier transform of output function. (*b*) Fourier transform of smearing function.

The photographs, Figs. 5-24 through 5-31, are oscilloscopic presentations of some of the operations described above. The displays were obtained by digital instrumentation techniques.

The previous example illustrated the use of a filter designed to maximize the signal-to-noise ratio of the resultant data. No attention was paid to signal fidelity or transient response. The sharp corners at $\pm s_0$ in the preceding example will cause a considerable amount of ringing on transients. Many other types of filter functions can be used, depending on the particular application. For example, a simple rectangular filter

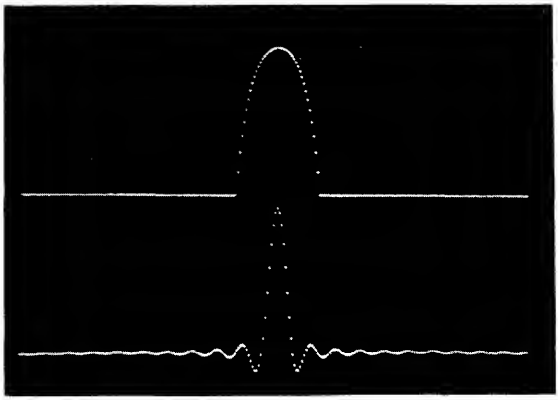


FIG 5-26 (*a*) Matched filter for optimizing quotient signal-to-noise ratio. (*b*) Impulse response of matched filter.

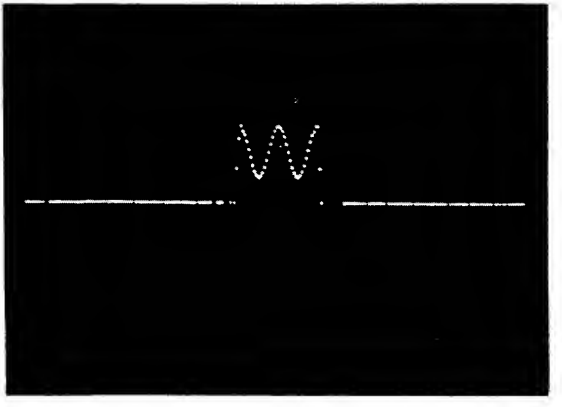


FIG 5-27 Real part of quotient of transforms of output and smearing functions.

might be most useful if the coherent signal is sharply band limited, or if transients are of no particular concern.

There is another type of filter that has some useful properties which will be described in more detail. Consider the following Fourier transform pair:

$$\cos\left[2\pi s_0(t^2 - t_0^2)^{1/2}\right] \leftrightarrow \tfrac{1}{2}[\delta(s + s_0) + \delta(s - s_0)] + \pi s_0 t_0 \frac{I_1[2\pi t_0(s_0^2 - s^2)^{1/2}]}{(s_0^2 - s^2)^{1/2}} \tag{5-3-93}$$

where $|s| \leq s_0$.

The time waveform has a peak amplitude of $\cosh(2\pi s_0 t_0)$ at $t = 0$ and a uniform ripple of unit amplitude for all $|t| \geq t_0$. The area is $\pi t_0 I_1(2\pi s_0 t_0)$.

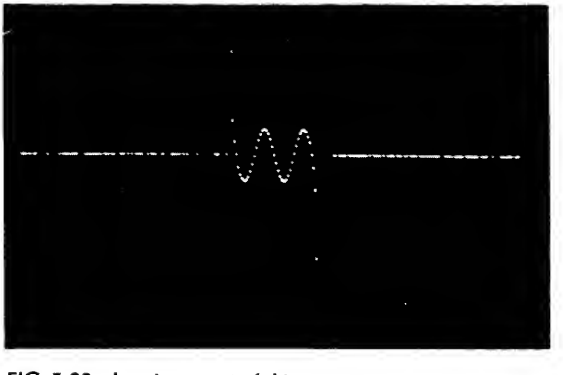


FIG 5-28 Imaginary part of this quotient.

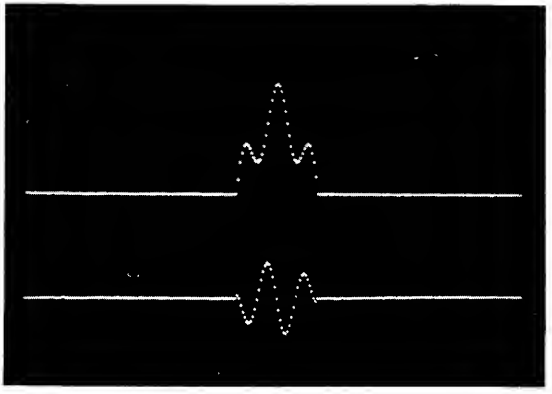


FIG 5-29 (*a*) Real part of quotient. (*b*) Imaginary part of quotient by after multiplication matched filter.

Thus convolution with this function will introduce a prescribed amount of ringing on transients. This time function minimizes the main-lobe width for a particular ripple amplitude. The resultant frequency filter is of finite extent, being zero for $|s| > s_0$. For reference, see Ref. 3, pairs 871.2 and 619.

Consider the previous example of gaussian smearing with a rectangular noise distribution on $G(s)$. What is the possible resolution improvement for a given amount of ripple and for some initial signal-to-noise ratio? There are various ways of measuring resolution, but the *equivalent width* of the time function will be used in this discussion. This is simply the

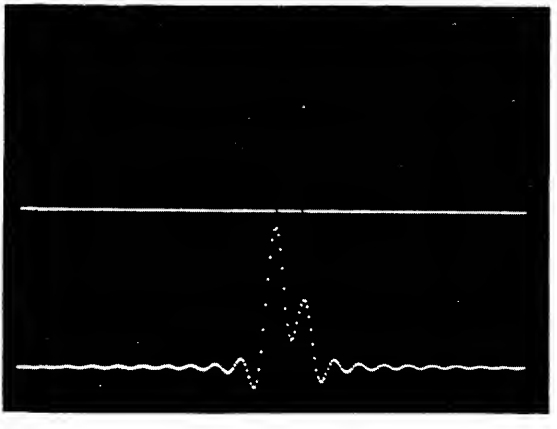


FIG 5-30 (*a*) Original pair of input delta functions. (*b*) Reconstructed input waveform.

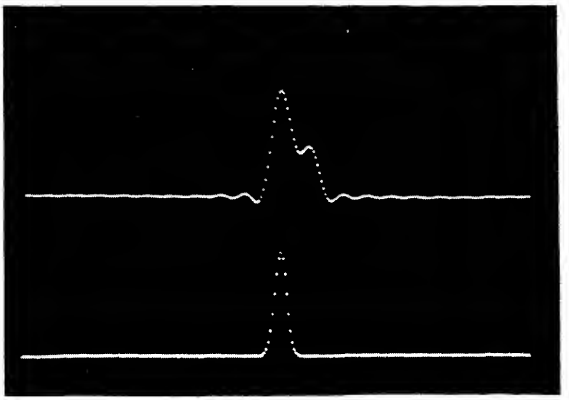


FIG 5-31 (*a*) Smoothed version of reconstructed input waveform. (*b*) Smoothing function with which output data was convolved.

width of a rectangle whose area and height are the same as the waveform in question. The equivalent width of the assumed gaussian smearing function is

$$\omega_g = \frac{1}{b(2\pi)^{1/2}} \tag{5-3-94}$$

Define a ripple factor r which is the ratio of peak ripple to main-lobe height in the time domain.

$$r = \operatorname{sech} 2\pi s_0 t_0$$

or

$$s_0 = \frac{1}{2\pi t_0} \cosh^{-1} \frac{1}{r} = b \left(2 \ln \frac{2k}{a} \right)^{1/2} \tag{5-3-95}$$

The equivalent width of the filter function in the time domain is

$$\omega_f = \pi t_0 \frac{I_1(2\pi s_0 t_0)}{\cosh 2\pi s_0 t_0} \tag{5-3-96}$$

The signal-to-noise ratio of the original function $G(s)$ is obtained by dividing the signal variance (k^2 at the peak) by the noise variance $a^2/12$. This ratio in decibels is

$$R = 10 \log \frac{12k^2}{a^2}$$

and

$$s_0 = b \left(\frac{\ln 10}{10} R - \ln 3 \right)^{1/2} \tag{5-3-97}$$

TABLE 5-4 Possible width reduction by inverse filtering.

	Signal-to-noise ratio, dB										
Ripple, dB	20	30	40	50	60	70	80	90	100	110	120
20.00	0.7871	0.6115	0.5175	0.4567	0.4133	0.3803	0.3541	0.3327	0.3147	0.2994	0.2861
25.00	0.8889	0.6906	0.5844	0.5157	0.4667	0.4295	0.3999	0.3757	0.3554	0.3381	0.3231
30.00	0.9787	0.7604	0.6435	0.5679	0.5139	0.4729	0.4403	0.4137	0.3914	0.3723	0.3558
35.00	1.0603	0.8238	0.6971	0.6153	0.5568	0.5123	0.4770	0.4482	0.4240	0.4033	0.3854
40.00	1.1358	0.8824	0.7476	0.6590	0.5964	0.5488	0.5110	0.4801	0.4542	0.4320	0.4129
45.00	1.2063	0.9372	0.7931	0.6999	0.6334	0.5828	0.5427	0.5099	0.4824	0.4589	0.4385
50.00	1.2728	0.9889	0.8368	0.7385	0.6683	0.6150	0.5726	0.5380	0.5090	0.4842	0.4627
55.00	1.3359	1.0379	0.8783	0.7752	0.7015	0.6455	0.6010	0.5647	0.5342	0.5082	0.4856
60.00	1.3962	1.0847	0.9179	0.8101	0.7331	0.6746	0.6281	0.5901	0.5583	0.5311	0.5075
65.00	1.4539	1.1296	0.9559	0.8436	0.7634	0.7025	0.6541	0.6145	0.5814	0.5531	0.5285
70.00	1.5094	1.1727	0.9924	0.8758	0.7926	0.7293	0.6791	0.6380	0.6036	0.5742	0.5487
75.00	1.5629	1.2143	1.0276	0.9069	0.8207	0.7552	0.7032	0.6606	0.6250	0.5945	0.5682
80.00	1.6147	1.2545	1.0616	0.9369	0.8478	0.7802	0.7264	0.6825	0.6457	0.6142	0.5870
85.00	1.6648	1.2934	1.0945	0.9660	0.8742	0.8044	0.7490	0.7037	0.6657	0.6333	0.6052
90.00	1.7134	1.3312	1.1265	0.9942	0.8997	0.8279	0.7709	0.7242	0.6852	0.6518	0.6229
95.00	1.7607	1.3679	1.1576	1.0217	0.9245	0.8507	0.7922	0.7442	0.7041	0.6698	0.6401
100.00	1.8068	1.4037	1.1879	1.0484	0.9487	0.8730	0.8129	0.7637	0.7225	0.6873	0.6568
105.00	1.8517	1.4386	1.2174	1.0744	0.9723	0.8947	0.8331	0.7827	0.7404	0.7044	0.6731
110.00	1.8955	1.4727	1.2462	1.0999	0.9953	0.9159	0.8528	0.8012	0.7580	0.7211	0.6891
115.00	1.9384	1.5060	1.2744	1.1247	1.0178	0.9366	0.8721	0.8193	0.7751	0.7374	0.7046
120.00	1.9803	1.5385	1.3020	1.1491	1.0398	0.9568	0.8909	0.8370	0.7919	0.7533	0.7199

Define a width-reduction (resolution-improvement) ratio as

$$\omega = \frac{\omega_f}{\omega_g} = \left(\frac{\pi}{2}\right)^{1/2} \frac{r(\cosh^{-1} 1/r) I_1(\cosh^{-1} 1/r)}{[(\ln 10/10)\, R - \ln 3]^{1/2}} \tag{5-3-98}$$

A tabulation of ω for various values of r and R is included (Table 5-4).

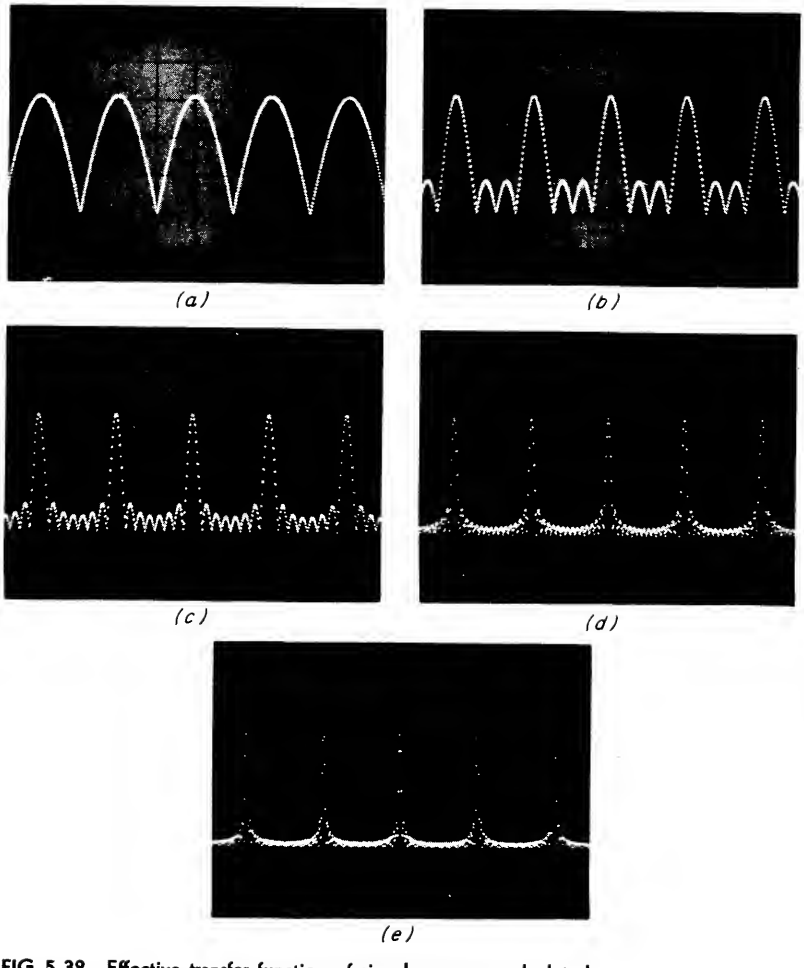


FIG 5-32 Effective transfer function of signal averager, calculated by digital computer. Averager acts like comb filter for repetitive waveforms. Since averager is timelocked to repetitive waveform, frequency components of waveform coincide exactly with centers of comb-filter teeth. Width of teeth is inversely proportional to number of repetitions averaged.

Note that the width reduction varies slowly with respect to both ripple and signal-to-noise ratio. Also see Fig. 5-32 for the computed transfer function of a signal averager.

5-4 Summary

Although the basic theory of Fourier transforms and statistics has been reviewed, the main emphasis in this chapter has been on various factors associated with the practical implementation of signal analysis. The use of digital techniques allows a considerable improvement in accuracy and a considerable increase in flexibility. Many of the errors in digital processing have been discussed. There are digitizing errors in signal amplitude, and there are errors caused by a finite integration time and by finite abscissa resolution. The implications of these errors in the frequency domain have been discussed. The importance of various averaging techniques has been considered along with a few words about convergence to a stable estimate. Filtering and inverse filtering (deconvolution) were introduced and the problems encountered are illustrated with a typical example. The concept of a matched filter has been discussed for maximizing signal-to-noise ratio. The neglect of these errors can introduce large anomalies in the results of digital signal analysis. It is therefore very important to know of the existence of these sources of error and to ascertain their effects before signal processing is completed.

It should be noted parenthetically that all the errors and limitations described above for digital signal processing have counterparts in signal analysis by analog means. Results are still affected by finite integration time and abscissa resolution, and the system still has noise. These limitations are imposed by nature and cannot be circumvented.

In the past few years, enough has been written on this general subject to fill several volumes, so a single chapter must, by necessity, be rather sketchy. There are many important topics concerning signal analysis by digital techniques which have not been mentioned here. For example, there is much more to be said about time windows and frequency line shapes. Conditions for the stability of a power spectrum estimate, and the variance of this estimate, are very important. The coherence function, along with the concepts of partial and multiple coherence, has not even been mentioned [12]. The importance of trend removal from the original time record prior to processing has been ignored. A technique of overlapping adjacent time records [13] is of fundamental importance in gleaning the maximum amount of information from a given time record. Digital filters are becoming very popular, and considerable literature can be found on this subject [14].

A considerable amount of work has been done on algorithms for per-

forming some of the basic operations such as the Fourier transform. Of course, there is a wide-open field in the study of nonstationary processes and techniques for measuring their characteristics.

References 15 to 17 are very useful for further information in this general field.

APPENDIX

Amplitude-Spectrum Averaging

Define two independent gaussian random variables x and y, each with the same variance σ^2. Consider x and y as the real and imaginary parts of a frequency spectrum. Define two new random variables

$$z = \sqrt{x^2 + y^2} \qquad 0 < z < \infty$$

$$\omega = \tan^{-1} \frac{y}{x} \qquad -\pi < \omega < \pi$$

where z represents the magnitude and ω represents the phase angle of a particular spectral point. Note that $x = z \cos \omega$ and $y = z \sin \omega$. The area element is $dA = z\, dz\, d\omega$. Thus the joint probability density is

$$p_{z\omega}(z,\omega) = \frac{z}{2\pi\sigma^2} e^{-(1/2\sigma^2)[z^2+\bar{x}^2+\bar{y}^2-2z\bar{x}\cos\omega-2z\bar{y}\sin\omega]}$$

The marginal probability densities are

$$p_z(z) = \int_{-\pi}^{\pi} p_{z\omega}(z,\omega)\, d\omega = \frac{z}{\sigma^2} I_0\left(\frac{z\sqrt{x^2+y^2}}{\sigma^2}\right) e^{-(1/2\sigma^2)(z^2+\bar{x}^2+\bar{y}^2)}$$

$$p_\omega(\omega) = \int_0^{\infty} p_{z\omega}(z,\omega)\, dz = \frac{1}{2\pi}[1 + \sqrt{\pi}\, u e^{u^2}(1 + \operatorname{erf} u)] e^{-(1/2\sigma^2)(\bar{x}^2+\bar{y}^2)}$$

where $u = (1/\sqrt{2}\,\sigma)(\bar{x}\cos\omega + \bar{y}\sin\omega)$, and $I_0(\xi)$ is the modified Bessel function of order zero. Note that $p_{z\omega}(z,\omega) \neq p_z(z)p_\omega(\omega)$, so z and ω are not independent. Since $p_\omega(\omega)$ is not needed in this discussion it will be ignored. For the special case when $\bar{x} = \bar{y} = 0$,

$$p_z(z) = \frac{z}{\sigma^2} e^{-\frac{1}{2}(z/\sigma)^2}$$

is a Rayleigh distribution, where $\bar{z} = \sigma\sqrt{\pi/2}$, and $\overline{z^2} = 2\sigma^2$. Thus the variance is $[2 - (\pi/2)]\sigma^2$. The moments for the general case are

$$\overline{z^m} = \int_0^{\infty} z^m p_z(z)\, dz$$

This integral is of the form [10]

$$\int_0^\infty t^{\mu-1} I_\nu(at) e^{-(pt)^2}\, dt = \frac{\Gamma[\frac{1}{2}(\mu+\nu)]}{2p^\mu \Gamma(\nu+1)} \left(\frac{a}{2p}\right)^\nu {}_1F_1\left(\frac{u}{2}, 1, \frac{a^2}{4p^2}\right)$$

where ${}_1F_1(a,b,\xi)$ is the confluent hypergeometric function [11], $\Gamma(\mu)$ is the gamma function, and $I_\nu(\xi)$ is the modified Bessel function of order ν.

$${}_1F_1(a,1,\xi) = \sum_{n=0}^{\infty} \frac{\Gamma(a+n)}{(n!)^2\Gamma(a)} \xi^n$$

Thus

$$\overline{z^m} = (\sqrt{2}\,\sigma)^m \Gamma(1+\tfrac{1}{2}m) e^{-(1/2\sigma^2)(\bar{x}^2+\bar{y}^2)} {}_1F_1\left(1+\tfrac{1}{2}m, 1, \frac{\bar{x}^2+\bar{y}^2}{2\sigma^2}\right)$$

Even-order moments can be expressed in elementary form:

$${}_1F_1(1,1,\xi) = e^\xi \quad \text{so} \quad \overline{z^0} = 1 \quad \text{area under } p_z(z)$$
$${}_1F_1(2,1,\xi) = (\xi+1)e^\xi \quad \text{so} \quad \overline{z^2} = \bar{x}^2 + \bar{y}^2 + 2\sigma^2$$

Unfortunately the calculation of the mean is not so easy.

$$\bar{z} = \sigma\sqrt{\frac{\pi}{2}}\, e^{-\xi} {}_1F_1\left(\frac{3}{2}, 1, \xi\right)$$

where

$$\xi = \frac{1}{2r^2} = \frac{\bar{x}^2+\bar{y}^2}{2\sigma^2}$$

For $\xi \to \infty$,

$${}_1F_1\left(\frac{3}{2}, 1, \xi\right) \to 2\sqrt{\frac{\xi}{\pi}}\, e^\xi \left(1+\frac{1}{4\xi}\right) \quad \text{and} \quad \bar{z} \to \sqrt{\bar{x}^2+\bar{y}^2+\sigma^2}$$

For $\xi = 0$,

$${}_1F_1\left(\frac{3}{2}, 1, \xi\right) = 1 \quad \text{and} \quad \bar{z} = \sigma\sqrt{\frac{\pi}{2}}$$

This first moment is readily calculated on a computer.

Analogous to the technique used in discussing power averaging, define a new random variable $\omega = z - z_0$, where z_0 is the random variable for noise alone ($\bar{x}^2 + \bar{y}^2 = 0$).

$$\bar{\omega} = \bar{z} - \bar{z}_0 = \sigma\sqrt{\frac{\pi}{2}} \left[e^{-\xi} {}_1F_1\left(\frac{3}{2}, 1, \xi\right) - 1\right]$$

$$n\sigma_\omega^2 = \overline{z^2} - \bar{z}^2 + \left(2 - \frac{\pi}{2}\right)\sigma^2 = \bar{x}^2 + \bar{y}^2 - \bar{z}^2 + \left(4 - \frac{\pi}{2}\right)\sigma^2$$

Let $\bar{\omega} = k\sigma_\omega$, where k is selected as before to achieve some probability of recognizing a coherent component against a noise background. For a prescribed value of k, it is useful to plot the required number of signals averaged, n, against r. Figure 5-33 illustrates this relationship for both power and amplitude averaging for various values of k. Note that n is proportional to k^2 so curves for other values of k are easy to construct. The relative efficiency of these two methods is apparent. Amplitude averaging requires a value of $n = (16/\pi) - 4 = 1.093$ times larger than

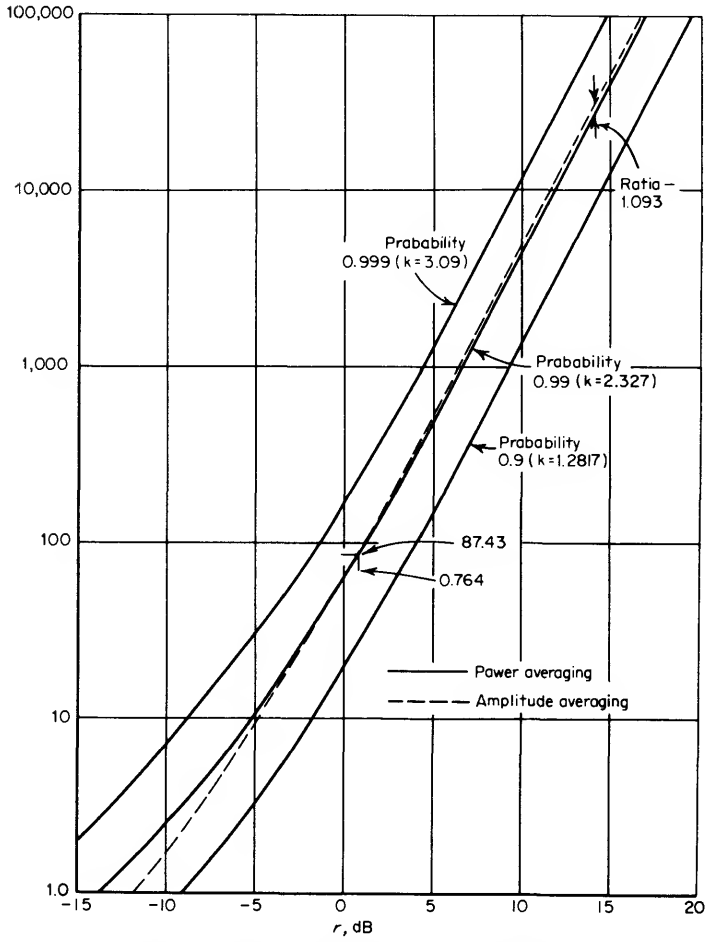


FIG 5-33 Required averaging (n) to resolve a coherent signal mixed with Gaussian noise within a prescribed probability (noise-to-signal ratio is r).

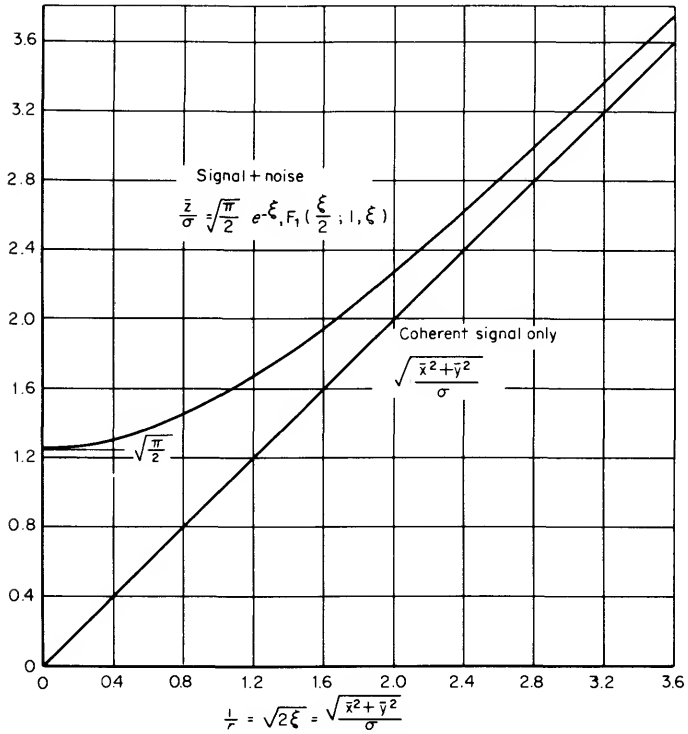


FIG 5-34 Amplitude of noisy spectral line compared to amplitude of coherent component.

power averaging for the same resolution probability as long as n is large. The two methods of averaging are equivalent for $r = 1.092$ (0.764 dB).

The interpretation of a power average is straightforward because the coherent signal amplitude is equal to the difference between a dominant spectral line and the noise base line. The interpretation of an amplitude average is more difficult and is best described graphically (see Fig. 5-34). The coherent amplitude is asymptotically proportional to the amplitude of a spectral line for small amounts of noise, but departs significantly from this proportionality as the noise increases. Thus it is difficult to compare directly the coherent amplitudes of signals mixed with noise. For these reasons the power averaging approach is generally preferred.

CITED REFERENCES

1. Bracewell, R. N.: "The Fourier Transform and Its Applications," McGraw-Hill Book Company, New York, 1965.
2. Papoulis, A.: "The Fourier Integral and Its Applications," McGraw-Hill Book Company, New York, 1962.

3. Campbell, G. A., and M. Foster: "Fourier Integrals for Practical Applications," D. Van Nostrand, 1942.
4. Cooley, J. W., and J. W. Tukey: An Algorithm for the Machine Calculation of Complex Fourier Series, *Math. Computation*, vol. 19, no. 90, pp. 297–301, 1965.
5. Gentleman, W. M., and G. Sande: "Fast Fourier Transforms—For Fun and Profit," 1966 Fall Joint Computer Conference, AFIPS Press, Montvale, N.J., 1969.
6. Papoulis, A.: "Probability, Random Variables, and Stochastic Processes," McGraw-Hill Book Company, New York, 1965.
7. Davenport, W. B., and W. L. Root: "Introduction to Random Signals and Noise," McGraw-Hill Book Company, New York, 1958.
8. Blackman, R. B., and J. W. Tukey, "The Measurement of Power Spectra," Dover Publications, Inc., New York, 1959.
9. Papoulis, A.: "Probability, Random Variables and Stochastic Processes," McGraw-Hill Book Company, New York, 1965, pp. 66–67.
10. Watson, G. N.: "A Treatise on the Theory of Bessel Functions," The Macmillan Company, New York, 1944.
11. Abramowitz, Milton, and I. A. Stegun: "Handbook of Mathematical Functions," U.S. Government Printing Office No. AMS-55, June, 1964.
12. Bendat and Piersol: "Measurement and Analysis of Random Data," John Wiley and Sons, Inc., New York, 1966, pp. 33, 103–122, 230–237.
13. Welch, P. D.: The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Short, Modified Periodograms, *IEEE Transactions on Audio and Electroacoustics*, vol. AU-15, pp. 70–72, June, 1967.
14. Gold and Radar: "Digital Processing of Signals," McGraw-Hill Book Company, New York, 1969.
15. *IEEE Transactions on Audio and Electroacoustics*, vol. AU-15, June, 1967.
16. *IEEE Transactions on Audio and Electroacoustics*, vol. AU-17, June, 1969.
17. *IEEE Transactions on Audio and Electroacoustics*, vol. AU-18, December, 1970.

CHAPTER SIX

FREQUENCY AND TIME MEASUREMENTS

Alan S. Bagley

Manager, Santa Clara Division
Hewlett-Packard Company, Santa Clara, California

The frequency of a repetitive signal, which is the number of cycles per unit of time, is one of the chief variables of interest in electrical communications, in the measurement of physical quantities, and in many natural phenomena. Very sophisticated instruments and techniques have been developed for precise frequency measurement. Frequency[1] and time are interdependent, and both are treated in this chapter. Some instruments are designed to measure both quantities, but many techniques are so specifically contrived to measure one or the other that these techniques are described separately here. There are sections dealing with the measurement of frequency, the period of a repetitive wave, and time interval on a one-shot basis.

Several chapters have emphasized the necessity for an accurate "standard of reference" in making an accurate measurement. The reference

[1] The unit of frequency, the *hertz* (abbreviated Hz), is of course a *derived* unit equivalent to one event per second.

standard seems particularly important in time and frequency measurements, along with the precise comparison of the unknown with the reference that is chosen. The development of good standards for both frequency and time interval—as well as historical or *epochal* time instants —has been an exciting adventure in natural philosophy.

6-1 Time Definitions and Standards

For many centuries, the rotation of the earth about its axis, viewed with respect to the sun, was used to set up a uniform time scale. Astronomers gradually increased the precision of their observations and found that the rotation of the earth was not really uniform. Even after they applied every correction known to them, they found unpredictable irregularities and long-term drifts. However, long after astronomers became aware of these variations (until 1956) the second was defined as $1/86{,}400$ part of an average rotation of the earth about its axis with the sun as the reference direction.

During the past two decades, largely through the coordinating efforts of The International Committee on Weights and Measures, time scale definitions have proliferated, although much of the basic work was done earlier. The brief statements below of the time definitions should be good background knowledge for the serious student.

The reader should remember the two separate problems present in timekeeping. One is the determination of the epoch, or the instant of time in history, and the other is the determination of time interval, which depends upon a satisfactory time unit, a standard second of time. The engineer's interest is usually focused on time interval, and the precision of definition of the second has increased about 10^5 fold during the past three decades, but it should be noted that the determination of the epoch, with respect to other fairly recent epochs, still is important to the engineer. A part of our defense systems depends upon this determination.

Ephemeris time is based upon the earth's orbital revolution around the sun. In 1956, The International Committee on Weights and Measures defined the unit of ephemeris time as follows: ". . . the second is the fraction 1/31,556,925.9747 of the tropical year for January 0, 1900 at 12 hours Ephemeris Time" [1]. (To the astronomer, that epochal instant occurs at noon on the first day of January.) A tropical year is the time interval, taken symmetrically about a given epoch required for the sun to increase in mean longitude by 360°, measured along the ecliptic from the vernal equinox.

The second of ephemeris time was an improvement over the second as previously defined. Theoretically it was invariable, but astronomers

could not realize, or establish, the definition with sufficient accuracy to make it serve the purpose of a universal day-to-day standard.

What was needed was an invariable time interval in nature that could be observed and measured by available electronic instruments with great precision. This requirement, coupled with recent complex development in the electronic arts, led to the use of transitions that occur spontaneously and continually in energy states in atoms. In 1964, using the atomic transitions, the International Committee adopted a new standard that is called *atomic time* and is based upon a specific hyperfine transition in cesium 133. The definition is:

'The standard to be employed is the transition between the two hyperfine levels F = 4, mF = 0 and F = 3, mF = 0 of the fundamental state $^2S_{½}$ of the atom of cesium 133 undisturbed by external fields and the value 9,192,631,770 hertz is assigned.'

Atomic time is far more readily observable than ephemeris time, for an electronic instrument with the capability of counting out the required number of transitions in cesium vapor can be set up anywhere on earth (or outer space). In a recent check of time scales kept by laboratories in six countries over a period of nearly 2 years, the maximum observed difference was about 100 μsec. Atomic *frequency* thus serves as a basis for a precise physical time scale.

While the atomic definition of the second has served well to give accurate, immediate time scales and intervals, some procedures, such as precise navigation and satellite tracking, require correlation with the rotation of the earth. Several other time-of-day scales have been devised as a result of these requirements. One of the first was *mean solar time*, based upon the average interval for all solar days during the year. The mean solar second is $\frac{1}{86,400}$ of a mean solar day. This definition avoids the day-to-day variations caused by the tilt of the earth's axis and by orbital eccentricity.

As corrections were progressively made for more and more of the *cyclical* irregularities in the earth's rotation, various versions of *universal time* evolved. One such scale is still in wide use, although it still has some unpredictable and secular variation.

Sidereal time is still another scale. Its day is based upon the rotation of the earth with respect to the stars rather than the sun. Strictly speaking, the sidereal day is the interval between successive transits of the first point of Aries over the upper meridian of any place, and it is about 23 h, 56 min, and 4.09 sec.

6-2 Standard Frequency and Time-Signal Broadcasts

There are government time observatories or laboratories in many of the major countries of the world. It is sensible, and indeed essential, to

maintain extremely accurate frequency standards in these laboratories. To make the work of the laboratories widely available is obviously desirable, and radio has been used since the early days of radio broadcasting to disseminate standard time and frequency information. These transmissions make possible the setting of secondary standard clocks and oscillators to an accuracy of a millisecond or thereabouts in time of day and an accuracy in frequency of 1 part in 10^7 to 10^{11}, depending upon carrier frequency, propagation path, and the kind of instrumentation used.

Frequency and time transmitters are divided into two classes, according to carrier frequency. The high-frequency stations broadcast at frequencies between 3 and 30 MHz, and the low to very low frequency stations mainly use the spectrum between about 10 and 75 kHz. A few entertainment stations, such as the one at Droitwich, England, have their carriers very stably regulated by government frequency standards. Further information on the signals available in the United States can be obtained from the National Bureau of Standards (Frequency and Time Broadcast Services), Boulder, Colorado 80301.

In addition to the regulated carrier frequencies, these special broadcast stations emit various combinations of standard audio modulations, pulses, "ticks," and breaks to aid in timekeeping and frequency comparison.

A *constant* propagation time could be taken into account in standard broadcasts, but all propagation times vary. The variations are especially troublesome in the high-frequency range, where the paths include one or more reflections from the ionosphere. As the reflecting layers drift up and down, a doppler-effect frequency shift is observed at the receiver, the magnitude of which depends upon the velocity of layer movement. Frequency shifts of the order of several parts in 10^7 can easily occur in high-frequency systems. The stability of very low frequency (VLF) signals is much better, since the propagation is mainly by ground wave.

Actually, one should think of two effects. One is the change in instantaneous frequency caused by doppler effect as the path length changes. The other is a variable time delay depending on different propagation path lengths.

A basic scheme for comparing a local frequency standard with the broadcasts from a master station is shown in Fig. 6-1. Observe that frequencies are not compared directly. Instead, the master station transmits a sharp pulse or tick of modulation once every second, or more specifically, after every nth cycle of the carrier frequency, where n is the nominal frequency in hertz. Similarly, a tick is derived from the local frequency standard to be studied; if the error in the local standard were zero, the ticks produced by it would be at intervals of exactly 1 sec. The comparison instrument can be an oscilloscope, the sweep being triggered by the local ticks and the master ticks being applied to the vertical amplifier.

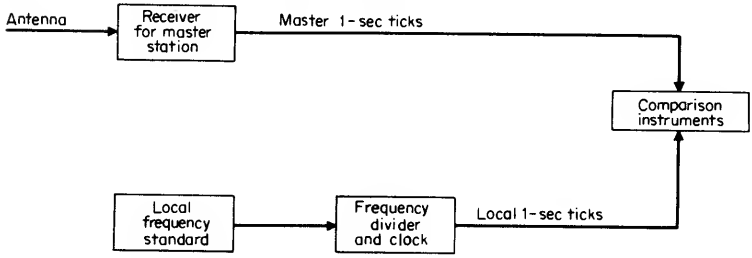


FIG 6-1 Frequency and time comparison system.

The advantage of the method in Fig. 6-1 is that extremely small errors in frequency can be measured if the relative drift in the timing of the two ticks is observed long enough. Of course, one measures only the *average* frequency by this method, and short-term stability must be determined separately.

In the VLF range, the fact that very low carrier frequencies cannot be modulated and demodulated accurately with sharp pulses leads to the use of phase comparisons with the VLF stations. Several methods are employed for phase comparison, but the block diagram in Fig. 6-2 shows one of the best schemes. This comparator provides for phase comparison between the 60-kHz signal from National Bureau of Standards station WWVB and a local frequency standard. Such comparisons serve for calibrating high-quality frequency standards or for monitoring

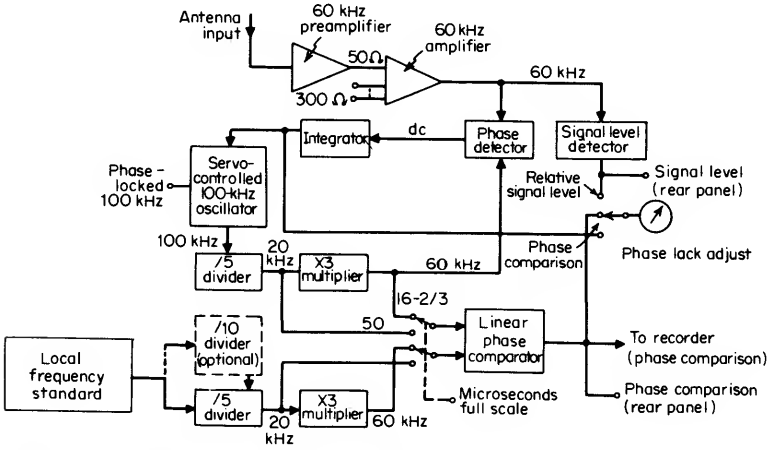


FIG 6-2 Simplified block diagram of a VLF comparator system. (Hewlett-Packard Co. Model 117A)

atomic-frequency standards. The VLF comparator thus provides a link between house frequency standards and the United States frequency standard.

The VLF comparator [2] in Fig. 6-2 is a complete system (exclusive of local standard). The instrument phase-tracks a voltage-controlled oscillator with WWVB. The local frequency standard is then compared with the phase-tracking oscillator. The comparator's strip-chart recorder makes a continuous recording of the phase difference, measured in microseconds.

Again refer to Fig. 6-2. For reasons that will be clear later, it is desirable to convert the amplified 60-kHz signal from WWVB into two phase-locked, constant-amplitude signals at 20 and 60 kHz. A closed loop is used to lock the phase of a 100-kHz oscillator, the output frequency of which is divided by 5 and multiplied by 3 to return to a 60-kHz signal, which is phase compared with the received signal. The dc output of the phase comparator is used to close the control loop.

The local frequency standard, which is normally at either 100 kHz or 1 MHz, is also converted to 20- and 60-kHz signals by multipliers and dividers. Then the signals derived from the local standard and the received standard transmission are fed to a linear phase comparator, which produces a voltage proportional to phase difference. This phase difference is either read on a meter or fed to a self-contained strip-chart recorder, or both. The time derivative of the recorded output (the slope of the trace) is proportional to frequency difference. If the two 20-kHz derived signals are phase compared, rather than the 60-kHz signals, the full-scale output of the linear phase comparator obviously denotes three times the *time* interval between master and local standards.

The phase-locked 100-kHz output serves as a convenient local-standard-frequency generator in itself, even when no local oscillator of good quality is available.

In the continental United States, frequency-standard comparisons to an accuracy of a part in 10^{10} can be approached in an 8-h period. A 24-h period may give 2 parts in 10^{11}, and a 30-day period may give accuracies of parts in 10^{12}. The local standard being calibrated must of course be of a stability commensurate with the realization of such high accuracies.

National Bureau of Standards station WWVB at Fort Collins, Colorado, is phase locked to the United States frequency standards and is kept to within a tolerance limit of $\pm 2 \times 10^{-11}$. The WWVB carrier frequency is referenced to the atomic second rather than to the second of universal time.

Since station WWVB is amplitude modulated with binary-coded time signals, the addition of an external strip-chart recorder to the phase-comparator instrument makes it possible to obtain a recording that tells

the day of the year, the hour, the minute, and the correction in milliseconds to arrive at UT2.

Detailed information on time and frequency comparison instruments can be obtained from the following sections and from various references [3,4].

6-3 Time and Frequency Standards

This section will present some of the theoretical and practical aspects of fluctuations in frequency standards and discuss the measurement of these fluctuations, or noise. The noise sources will be identified and evaluated. While this material may appear to be highly specialized for the electronics engineer, its comprehension at a technical and quantitative level is essential if the engineer becomes involved in complex communications systems, precise navigation, space systems, or even the intelligent use of time and frequency instrumentation.

Some of the design requirements for frequency standard instruments are also presented.

The following analytical treatment is taken directly from a paper by Cutler [5].

Analytical Treatment. The signal from an oscillator may be described by

$$f(t) = A(t) \cos [\omega_0 t + \Phi(t)] \tag{6-3-1}$$

where $f(t)$ represents a voltage or current, $A(t)$ and $\Phi(t)$ are slowly varying functions of time, and ω_0 is a constant. $A(t)$ is the amplitude of the signal and is assumed not to contribute to frequency fluctuations. The time origin and ω_0 are chosen so that $\Phi(t)$ has zero time average and $|\Phi(t)| \leq C < \infty$ for all time t, where C is some positive constant. These conditions simplify the mathematics (but will have to be relaxed later). The instantaneous angular frequency is

$$\omega(t) = \frac{d}{dt}[\omega_0 t + \Phi(t)] = \omega_0 + \dot{\Phi}(t) \tag{6-3-2}$$

In all that follows we shall refer to angular frequency as frequency. The average frequency is

$$\begin{aligned} \langle \omega(t) \rangle &= \lim_{T \to \infty} T^{-1} \int_{-T/2}^{T/2} \omega(t)\, dt \\ &= \omega_0 + \lim_{T \to \infty} \frac{\Phi(T/2) - \Phi(-T/2)}{T} \\ &= \omega_0 \end{aligned} \tag{6-3-3}$$

Therefore, $\Phi(t)$ is the instantaneous phase angle of the oscillator with

respect to an ideal oscillator of frequency ω_0. Let $\dot{\Phi}(t) = \Omega(t)$. The frequency departure averaged over time τ is

$$\langle \Omega_\tau(t) \rangle = \langle \dot{\Phi}_\tau(t) \rangle = \tau^{-1} \int_{t-\tau/2}^{t+\tau/2} \dot{\Phi}(t')\, dt'$$
$$= \tau^{-1} \left[\Phi\left(t + \frac{\tau}{2}\right) - \Phi\left(t - \frac{\tau}{2}\right) \right] \tag{6-3-4}$$

where $\langle \Omega \rangle$ signifies the average (time or statistical) of Ω and $\langle \Omega_\tau(t) \rangle$ signifies the finite time average at time t,

$$\langle \Omega_\tau(t) \rangle = \tau^{-1} \int_{t-\tau/2}^{t+\tau/2} \Omega(t')\, dt' \tag{6-3-5}$$

The phase averaged over time τ is

$$\langle \Phi_\tau(t) \rangle = \tau^{-1} \int_{t-\tau/2}^{t+\tau/2} \Phi(t')\, dt' \tag{6-3-6}$$

The phase difference over time τ is

$$\Delta\Phi_\tau(t) = \Phi\left(t + \frac{\tau}{2}\right) - \Phi\left(t - \frac{\tau}{2}\right) = \langle \Omega_\tau(t)\tau \rangle$$

Now consider that

$$R_\Phi(\tau) = \left\langle \Phi\left(t + \frac{\tau}{2}\right) \Phi\left(t - \frac{\tau}{2}\right) \right\rangle$$
$$= \lim T^{-1} \int_{-T/2}^{T/2} \Phi\left(t + \frac{\tau}{2}\right) \Phi\left(t - \frac{\tau}{2}\right) dt \tag{6-3-7}$$

is the autocorrelation function of the phase. Similarly, $R_\Omega(\tau)$ is the autocorrelation function of the frequency departure. Writing these both as functions of τ only implies that Φ and Ω are stationary in the wide sense [6].

$$S_\Phi(\omega) = \int_{-\infty}^{\infty} R_\Phi(\tau) \exp(-i\omega\tau)\, d\tau$$
$$= 2 \int_0^{\infty} R_\Phi(\tau) \cos \omega\tau\, d\tau \tag{6-3-8}$$

$$R_\Phi(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_\Phi(\omega) \exp(i\omega\tau)\, d\omega$$
$$= \pi^{-1} \int_0^{\infty} S_\Phi(\omega) \cos \omega\tau\, d\omega \tag{6-3-9}$$

so that $S_\Phi(\omega)$ and $R_\Phi(\tau)$ are Fourier transforms of each other [7], where $S_\Phi(\omega)$ is the power spectral density of the phase (we use the two-sided power spectrum). In the same way $R_\Omega(\tau)$ and $S_\Omega(\omega)$ are Fourier transforms of each other, where $S_\Omega(\omega)$ is the power spectral density of the frequency departure.

A useful measure of fluctuation is the standard deviation σ.

$$\sigma(X) = \langle (X - \langle X \rangle)^2 \rangle^{1/2} = (\langle X^2 \rangle - \langle X \rangle^2)^{1/2} \tag{6-3-10}$$

The standard deviation of the various quantities defined earlier can be written in terms of the autocorrelation functions:

Standard deviation of average phase,

$$\sigma\langle \Phi_\tau(t) \rangle = \left[\frac{2}{\tau} \int_0^\tau R_\Phi(\tau') \left(1 - \frac{\tau'}{\tau}\right) d\tau' \right]^{1/2} \tag{6-3-11}$$

Standard deviation of phase difference,

$$\sigma[\Delta\Phi_\tau(t)] = \{2[R_\Phi(0) - R_\Phi(\tau)]\}^{1/2} \tag{6-3-12}$$

Standard deviation of average frequency,

$$\sigma\langle \Omega_\tau(t) \rangle = \tau^{-1}\{2[R_\Phi(0) - R_\Phi(\tau)]\}^{1/2} \tag{6-3-13}$$

Standard deviation of average frequency departure,

$$\sigma \frac{\langle \Omega_\tau(t) \rangle}{\omega_0} = \frac{1}{\omega_0 \tau} \{2[R_\Phi(0) - R_\Phi(\tau)]\}^{1/2} \tag{6-3-14}$$

The last two may equally well be written in terms of $R_\Omega(\tau)$:

$$\sigma\langle \Omega_\tau(t) \rangle = \left[\frac{2}{\tau} \int_0^\tau R_\Omega(\tau') \left(1 - \frac{\tau'}{\tau}\right) d\tau' \right]^{1/2} \tag{6-3-15}$$

$$\sigma \frac{\langle \Omega_\tau(t) \rangle}{\omega_0} = \frac{1}{\omega_0} \left[\frac{2}{\tau} \int_0^\tau R_\Omega(\tau') \left(1 - \frac{\tau'}{\tau}\right) d\tau' \right]^{1/2} \tag{6-3-16}$$

Also, we have:

Standard deviation of phase,

$$\sigma[\Phi(t)] = [R_\Phi(0)]^{1/2} = \left[\pi^{-1} \int_0^\infty S_\Phi(\omega)\, d\omega \right]^{1/2} \tag{6-3-17}$$

Standard deviation of frequency,

$$\begin{aligned} \sigma[\Omega(t)] = [R_\Omega(0)]^{1/2} &= \left[\pi^{-1} \int_0^\infty S_\Omega(\omega)\, d\omega \right]^{1/2} \\ &= \left[\pi^{-1} \int_0^\infty \omega^2 S_\Phi(\omega)\, d\omega \right]^{1/2} \end{aligned} \tag{6-3-18}$$

The preceding formulas hold for wide-sense stationary random processes or for time functions that have stationary means and autocorrelation functions that depend only on the time difference τ.

Measurement Techniques. A general system for making measurements of some of the standard deviations described above is shown in Fig. 6-3. Two signal sources, slightly offset in average frequency, feed two identical

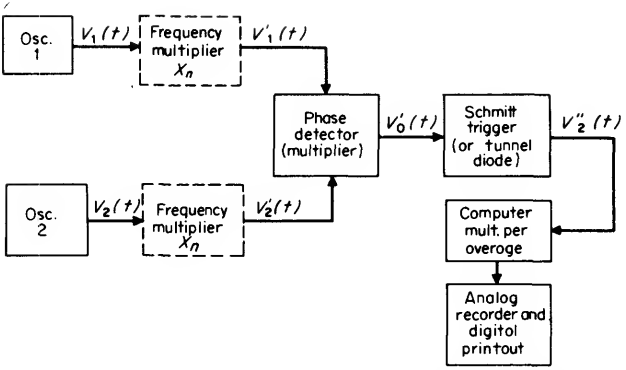


FIG 6-3 Multiple period measuring system for oscillator deviations.

channels through optional frequency multipliers to a phase detector. The difference frequency contains all the phase information and is used to trigger the Schmitt trigger at the zero crossings. The period (or multiple period) of the sharp leading edge of the Schmitt trigger is measured by the counter and displayed by the analog recorder, and each measurement is printed out on a digital recorder.

The theory behind this technique is as follows: Oscillator 1 has output

$$V_1(t) = A_1(t) \cos [\omega_1 t + \Phi_1(t)]$$

Similarly, oscillator 2 has output

$$V_2(t) = A_2(t) \cos [\omega_2 t + \Phi_2(t)] \tag{6-3-19}$$

In the frequency multipliers the amplitude changes get removed by limiting processes (if we have good design, there can be little conversion of the amplitude changes to phase changes). After multiplication, the signals are

$$\begin{aligned} V_1'(t) &\approx A_1 \cos [n\omega_1 t + n\Phi_1(t)] \\ V_2'(t) &\approx A_2 \cos [n\omega_2 t + n\Phi_2(t)] \end{aligned} \tag{6-3-20}$$

It is well known that both the instantaneous phase and the average frequency are multiplied by the factor n (provided the multiplier has sufficient bandwidth to encompass the full spectrum of the nth harmonic). The phase detector behaves as a multiplier. Its output is

$$\begin{aligned} V_0(t) \approx V_1'(t)V_2'(t) &= \tfrac{1}{2}A_1A_2 \cos \{n(\omega_1 + \omega_2)t + n[\Phi_1(t) + \Phi_2(t)]\} \\ &+ \tfrac{1}{2}A_1A_2 \cos \{n(\omega_1 - \omega_2)t + n[\Phi_1(t) - \Phi_2(t)]\} \end{aligned} \tag{6-3-21}$$

The sum frequency is filtered out, which leaves only the difference frequency term. If the two signal sources have exactly the same statistics

for $\Phi_1(t)$ and $\Phi_2(t)$ but are uncorrelated, then all the fluctuation can be assumed to be in one channel and to be $\sqrt{2}$ times as large as that channel alone, while the other channel can be assumed to be perfect. Let

$$\omega_1 - \omega_2 = \Delta\omega \tag{6-3-22}$$

Then the signal which feeds the Schmitt trigger is

$$V_0'(t) = \tfrac{1}{2}A_1A_2 \cos [n\,\Delta\omega t + n\Phi(t)] \tag{6-3-23}$$

where $\Phi(t) = \Phi_1(t) - \Phi_2(t)$. The Schmitt trigger gives a sharp pulse out each time the signal crosses zero going in, say, the negative direction. This occurs for t such that

$$n\,\Delta\omega t + n\Phi(t) = \tfrac{1}{2}\pi + 2\pi M$$

where M is any integer. Suppose the counter is set to count N periods and the gate opens at t_0 such that

$$n\,\Delta\omega t_0 + n\Phi(t_0) = \tfrac{1}{2}\pi$$

The gate will close at $t_0 + \tau$ to make

$$n\,\Delta\omega(t_0 + \tau) + n\Phi(t_0 + \tau) = \tfrac{1}{2}\pi + 2\pi N$$

Subtracting the first from the second, we get

$$n[\Delta\omega\tau + \Phi(t_0 + \tau) - \Phi(t_0)] = 2\pi N \tag{6-3-24}$$

Let

$$\tau = \frac{2\pi N}{n\,\Delta\omega} - \Delta\tau \equiv \tau_0 - \Delta\tau$$

$$\tau_0 = \frac{2\pi N}{n\,\Delta\omega} \tag{6-3-25}$$

Then

$$\Phi(t_0 + \tau) - \Phi(t_0) = \Delta\omega\,\Delta\tau \tag{6-3-26}$$

Since τ is not constant, the time difference $\Delta\tau$ between successive measurements is not constant, but if $\Delta\omega\,\Delta\tau \ll 1$ and $\dot{\Phi}(t_0)\,\Delta\tau \ll 1$, then only very small error is caused by replacing $\Phi(t_0 + \tau)$ with $\Phi(t_0 + \tau_0)$. The process of averaging over many measurements helps here. The multiple-period technique thus measures essentially $\Phi(t + \tau_0) - \Phi(t) = \Delta\Phi_{\tau_0}(t)$. Therefore,

$$\Delta\Phi_{\tau_0}(t) \approx \Delta\omega\,\Delta\tau \tag{6-3-27}$$

Almost any desired averaging time τ can be obtained by varying N, n, or $\Delta\omega$. By making many measurements of τ in succession, the standard

deviation can be estimated as

$$\sigma[\Delta\Phi_{\tau_0}(t)] \approx \Delta\omega\, \sigma(\Delta\tau)$$
$$\approx \Delta\omega \left[m^{-1} \sum_{i=1}^{m} \tau_i^2 - \left(m^{-1} \sum_{i=1}^{m} \tau_i \right)^2 \right]^{1/2} \tag{6-3-28}$$

where τ_i is the ith measurement and m is the total number of measurements, which should be large (about 100) to give a good estimate. It is wise to remove the drift during the observation time by subtracting the best straight line based on a least-squares fit from the data. For $m = 100$, this leads to

$$\sigma[\Delta\Phi_{\tau_0}(t)] = \Delta\omega \left\{ \frac{1}{99.99 \times 10^6} \left[999{,}900 \sum_{i=1}^{100} \tau_i^2 - 40{,}602 \left(\sum_{i=1}^{100} \tau_i \right)^2 \right.\right.$$
$$\left.\left. - 12 \left(\sum_{i=1}^{100} \tau_i i \right)^2 + 1{,}212 \left(\sum_{i=1}^{100} \tau_i i \right) \left(\sum_{i=1}^{100} \tau_i \right) \right] \right\}^{1/2} \tag{6-3-29}$$

The order of the data must be preserved for this formula. The other quantities of interest, such as Eqs. (6-3-13) and (6-3-14), may be estimated from Eq. (6-3-29) in an obvious way. The subtraction of the best straight line corresponds to filtering out the low-frequency fluctuations.

Equation (6-3-29) is difficult to apply, but it is given for information. The subtraction of the best straight line for the data involves the substitution of a variable $\tau_i' = \tau_i - (a + bi)$ and the determination of the constants a and b to minimize the variance. For a more detailed discussion and a good bibliography, see Ref. 17.

Figure 6-4 shows a block diagram of a versatile system which allows the two oscillators to have zero offset. This feature allows the system noise to be evaluated by feeding both channels from one source. The offset is obtained by the frequency synthesizer, whose fluctuations do not degrade the measurement much since the oscillator fluctuations have been multiplied by 1,840 times in the 20-MHz difference frequency before the comparison is made. Figure 6-5 shows some typical results obtained with this system.

From a practical standpoint the multiple-period system gives good results over a range of τ from about 10^{-2} sec upward.

Quartz Frequency Standards. The piezoelectric effect in certain crystals, particularly quartz, has long been used to stabilize the frequencies of oscillators. A mechanical force applied properly to a quartz crystal produces an electrical charge; an alternating potential across the crystal produces mechanical motion. This reciprocal relationship allows a sharp mechanical resonance in a crystal to be viewed as if it were a high-Q electrical resonance when two electrodes are mounted on or near the crystal

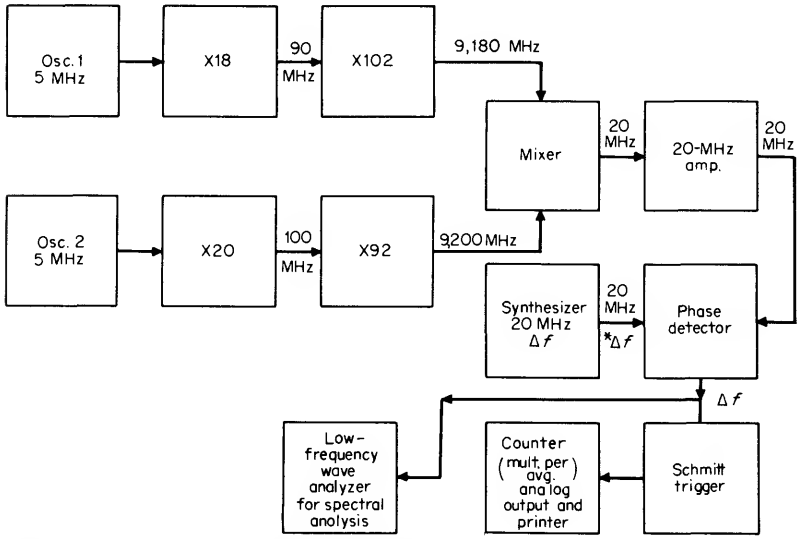


FIG 6-4 Versatile multiple-period measuring system.

surfaces. Limited space allows only a brief discussion of quartz-crystal technology, and the reader who wishes to study further should consult specialized references [8, 9, 10].

There are many forms in which quartz can be cut and mounted to serve as a resonator. The objectives are to obtain high Q, stability of resonant frequency versus age and temperature variation, reduction of unwanted oscillation modes, and freedom from the effects of mechanical shock and vibration. Even after the quartz has been well cut and properly mounted, it is necessary to enclose the crystal in a temperature-controlled oven to achieve good performance as a frequency standard. Very elaborate ovens with continuous feedback temperature control have been developed for the purpose.

For greatest frequency stability, the design of the oscillator goes beyond consideration of the crystal and its holder. Figure 6-6 is a representation of a quartz resonator connected in a Pierce oscillator configuration, a circuit that is frequently employed. Since the resonator appears electrically as a series resonant circuit of extremely high Q, shunted by C_0 (the total capacitance of holder and electrodes), either the pole or the zero lying near the mechanical resonant frequency can be used to determine the operating frequency of a suitable oscillator.

Every impedance shown in Fig. 6-6 affects the oscillator frequency slightly, even though the resonant Q is of the order of 10^6 to 10^7. The

oscillating frequency is simply the one at which the phase shift around the feedback loop is zero. Even though the phase changes extremely rapidly with frequency in the vicinity of mechanical resonance, variations in C_0, C_3, C_4, the input and output impedances of the amplifier, and small changes in phase shift through the amplifier all produce small changes in oscillator frequency. Therefore, all components in the circuit must be carefully selected for long- and short-term stability, and both Z_i and Z_o should be made as high as possible.

The frequency of oscillation varies slightly with the driving power into the crystal. It is best to control driving power automatically to about 1 μW, and *variations* in driving power are typically held to less than 0.01 percent. Obviously, power-supply voltages must be regulated to avoid fluctuations in driving power or spurious phase shifts in the active circuits.

To be complete, the circuit in Fig. 6-6 should show the output circuitry (that unavoidably loads the oscillator) and stray feedback from output circuits to the oscillator proper. Since in some cases present commercial oscillators achieve stabilities better than $\pm 5 \times 10^{-11}$ per 24 h, it should not be surprising that extreme attention is given to these seemingly minor factors. The design of modern electronic instruments is as demanding

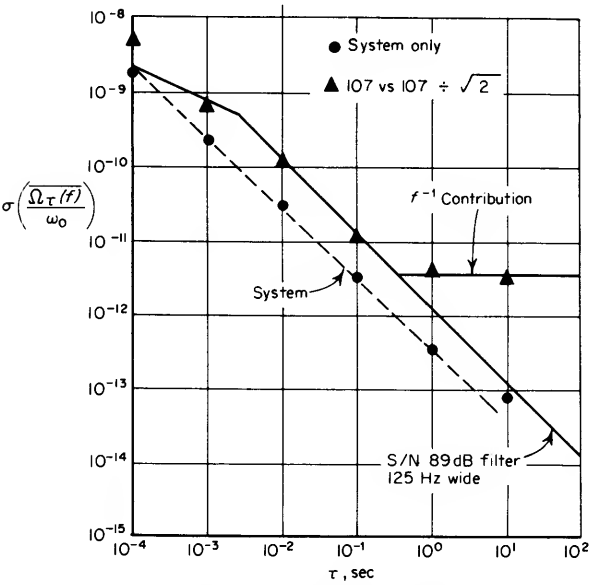


FIG 6-5 Measurements with multiple-period system.

of creativity, care, and theoretical considerations as any design problem in electronics.

Atomic Frequency Standards. The reader may not have studied quantum mechanics, upon which atomic frequency standards are technically based, but the presentation here will require only a few elementary principles. The chapter by Alan Bagley on Frequency and Time Measurements in Vol. 23 of the "Handbuch der Physik," edited by S. Flügge-Freiburg [11], gives a more thorough presentation, still not requiring advanced knowledge of quantum theory.

The energy levels in an atom can assume only a limited number of discrete levels, and transitions between two levels either give up or require the added energy difference between those two particular levels. Furthermore, there is a unique electromagnetic frequency related to each transition. The frequency is determined by the familiar relationship $E = h\nu$, where E is the difference in energy between the two states, h is Planck's constant and ν is the frequency. In other words, if energy at frequency ν is applied to an atom when the above relationship is satisfied, one quantum of that energy has the capability of effecting a sudden shift in energy level.

In a passive atomic resonator, transitions are induced by driving the atomic device with a signal derived from a quartz oscillator, which in turn is stabilized to produce the proper driving frequency. The active resonators are microwave masers, such as atomic hydrogen masers, which produce stimulated emission in the microwave range.

The principal passive standard (and the present basis for the United States frequency standard) is the cesium-beam resonator. For this standard, the quantum effects of interest arise in the nuclear magnetic hyperfine ground state of the atoms. A particularly appropriate transition occurs between the ($F = 4$, $m_F = 0$) and ($F = 3$, $m_F = 0$) hyperfine levels in the cesium-133 atom, arising from electron-spin–nuclear-spin

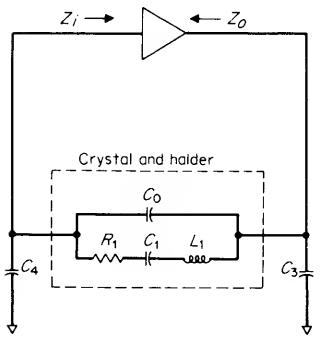


FIG 6-6 Equivalent circuit of a quartz crystal in an oscillator.

interaction [11]. This transition is appropriate for frequency control by reason of its relative insensitivity to external influences such as electric and magnetic fields and of its convenient frequency (in the microwave range, 9,192+ MHz).

Figure 6-7 is a simplified sketch of the cesium tube. Cesium atoms having all the permissible quantum-energy levels are slowly evaporated from an oven and collimated by slits into a beam. The state selector magnet A, with its vertical field, deflects each atom through an angle that depends upon the energy level of that atom. Only those atoms having the so-called F = 4, $m_F = 0$ state are permitted to form a beam through the reaction cavity of the tube, where a microwave magnetic field having a frequency of 9,192.631770 MHz has exactly the correct energy per quantum to flip the state of the atoms to F = 3, $m_F = 0$. Actually only a portion of the atoms are converted, but enough so that they can be sorted by the magnet E and made to impinge upon a hot wire that ionizes them. The resulting free electrons strike the first dynode of an electron multiplier and produce a dc component of output current.

If one plots output current versus excitation frequency, a very sharp resonant peak is found having a Q as high as 2×10^8. Figure 6-8 shows in a very simplified way how a crystal oscillator is controlled to produce a submultiple of the required excitation frequency. The driving signal, at 9,192+ MHz, is phase modulated by means of an audio oscillator. If the driving frequency is exactly at the resonant peak of the cesium tube, the output I_b contains only harmonics of the audio frequency and none of the fundamental. The phase-sensitive detector has a dc output proportional to the departure of the driver from the desired frequency, and the polarity of this error signal indicates whether the drive frequency is too low or too high. After high-frequency components in the error signal are removed by an integrating circuit, the signal is used to control and correct the operating frequency of a quartz oscillator running at about 5 MHz. A small, steady magnetic C field throughout the whole cavity area has been found to limit the atomic level transitions to the correct ones. Of course, the time constant of the integrating circuit also determines the

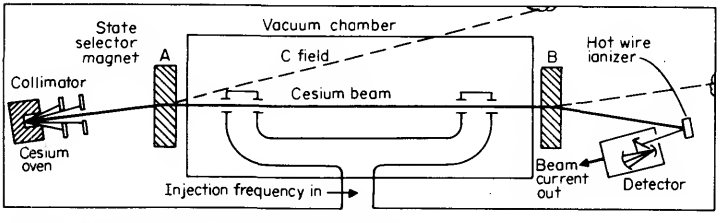


FIG 6-7 Schematic of cesium beam resonator.

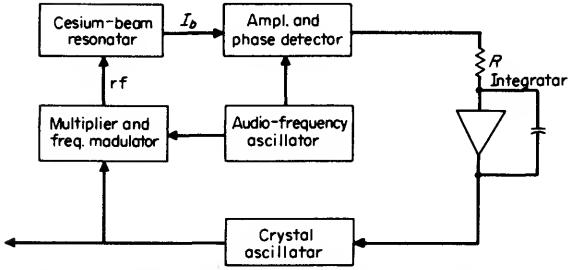


FIG 6-8 Control scheme for a cesium beam resonator.

bandwidth over which the quartz flywheel, with its good short-term stability, is controlled.

The cesium-beam resonator exhibits excellent long-term stability. A good portable commercial instrument (Hewlett-Packard model 5061A) has a specified accuracy of $\pm 1 \times 10^{-11}$, and a laboratory model at the National Bureau of Standards is even better. Atomic hydrogen masers produce less high-frequency noise in their outputs and are sometimes used when a spectrally pure signal is required. Other gas cells, such as ammonia and rubidium, have also been tried for masers and have some merits as frequency standards [12, 13]. Rubidium is also used in a passive gas cell with optical pumping.

6-4 *Frequency Measuring Instruments*

In the exciting history of electronic instruments, two types that have become commonplace are oscilloscopes (Chap. 11) and digital instruments based upon circuits that count pulses.

These counter circuits are constructed of devices or subcircuits that are n stable, usually bistable. If bistable circuits (flip-flops) are connected in tandem, a binary counter is formed, as in Fig. 6-9. Each flip-flop in this case is toggled from one output state to the opposite one when a

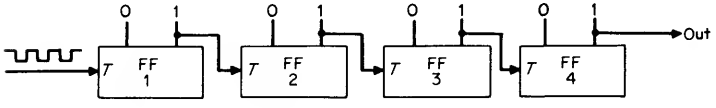


FIG 6-9 Basic binary counter diagram.

negative-going pulse is applied to the T input. Assuming that the state-indicating outputs 0 and 1 are most positive when "on," it follows that each flip-flop is toggled when the state of the previous one flips from 1 to 0. If all states are 0 initially, 16 negative input pulses are required to pro-

duce 1 negative output pulse, or in other words a scale-of-16 counter results.

However, a *decade* counter is easily made by feeding pulses between certain flip-flops in a binary counter with a scale of 16, as in Fig. 6-9, in addition to the pulse signal paths shown in that figure. This causes the flip-flops to change state in a more complicated sequence than in the straight binary fashion. A pulse signal emerges from FF4 after every tenth input pulse, and the total count stored in the unit at any given time is determined by the combination of conduction states among the flip-flops. By means of coincidence circuitry, the stored count operates a visual digital display. The whole assembly is called a *decade-counting assembly.*

A detailed treatment of the circuitry in a modern solid-state electronic counter is available in the service manuals of commercial counters [14].

Obviously, several decade-counter circuits can be put in tandem to produce a decimal counter. With two additional components, the manually controlled totalizing counter in Fig. 6-10 is formed. One of the additions is a gate circuit that either transmits pulses or prevents their transmission, depending upon the application of an electrical start-stop signal. This main gate is usually an arrangement of diodes or transistors very familiar to computer-circuit designers. The pulses, waves, or electrical "events" to be counted are fed to an amplifier and a trigger circuit that produces pulses of uniform and optimum shape, one pulse for each cycle or event of input. The counter counts pulses as long as the main gate is opened, and the total count is stored and displayed on the instrument panel until the counter circuits are reset electrically to zero. This is a very simple arrangement, and as additional components are added, an extremely versatile instrument emerges.

The most important additions are:

1. *An internal time-base oscillator having good frequency stability.* In simple, inexpensive instruments the time-base signal can be the ac power-

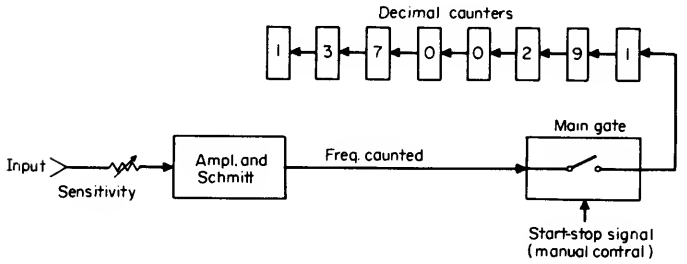


FIG 6-10 Manually controlled totalizing counter.

line voltage, but generally a quartz-crystal oscillator is used. The required quality of the quartz resonator and the sophistication of the constant-temperature oven and associated electronic circuitry depend upon the target specifications of the instrument. Short-term stabilities of better than 5 parts in 10^{11} and aging rates better than 5 parts in 10^{10} per day are available in commercial counters. The exact frequency of time-base oscillators can be set to an external house-frequency standard in most instruments of quality.

2. *Counting circuits that divide the crystal-oscillator frequency by powers of* 10. When the decade divider is connected into the system as in Fig. 6-11, along with a trigger circuit to develop pulses with optimum shape for fast, accurate control of the main gate, the gate can be opened for very accurate intervals of time that can be set in decade steps. In the figure, if the oscillator frequency is 1 MHz, the gate can be opened for precisely 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, or 10 sec.

Frequency Measurements. Figure 6-11 is a simplified block diagram of an electronic counter with its function switch set to measure the frequency of the input signal.

The input signal is first supplied to a signal shaper which converts the input signal (CW or pulses) to uniform pulses. The output of the shaper is then routed to decade-counting assemblies through a gate controlled by the counter's time base as shown in Fig. 6-11. The number of pulses totaled in the decade-counting assemblies for the selected period of time represents the frequency of the input signal. The frequency counted is displayed on a visual numerical readout, with a positioned decimal point, and is retained until a new sample is taken. The sample-rate control determines the display time of the frequency measurement being made and initiates the counter reset and the next measurement cycle.

The time-base selector switch selects the gating interval, positions the

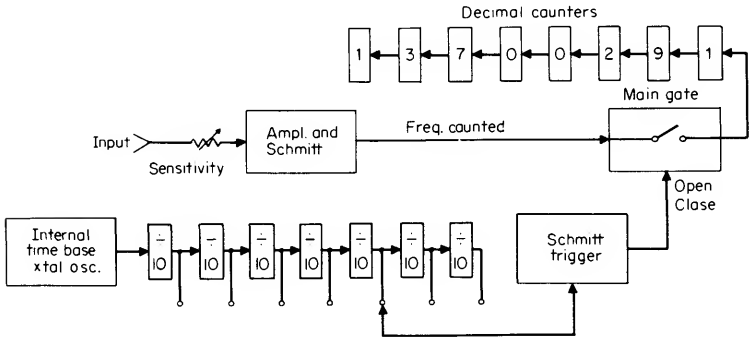


FIG 6-11 Basic electronic frequency counter.

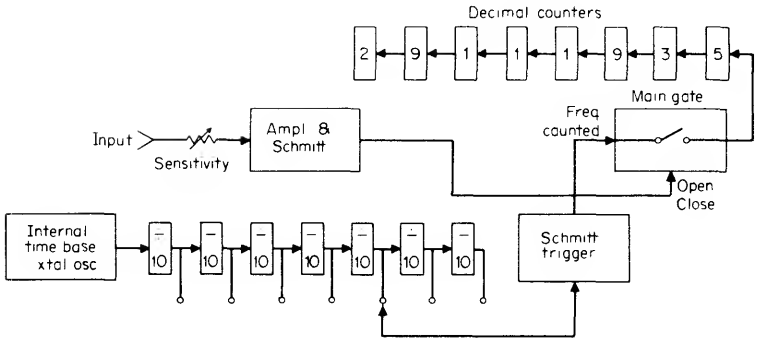


FIG 6-12 Function switch set to "Period."

decimal point, and selects the appropriate measurement units. For simplicity, the last two switching actions are not shown in the figure.

In the event that frequency measurement of low-level signals (down to 1 mV rms) is desired, a wideband amplifier can be placed ahead of the input terminals in Fig. 6-11. In some counters, such as the Hewlett-Packard model 5245, preamplifier plug-ins are readily available. A front-panel meter indicates whether the input level is adequate for the measurement.

Since the gate may open at any electrical angle of the frequency being measured and close at some other arbitrary angle, there is always an ambiguity of plus-or-minus one count or cycle. This is a fundamental limit on accuracy.

Period Measurements. Period $P = 1/f$, where f is frequency; therefore, period measurements are made with the counter functions arranged as shown in Fig. 6-12. The unknown input signal controls the main gate time, and the time-base frequency is counted in the decade-counting assemblies. The input-shaping circuit commonly selects the positive-going zero axis crossing of successive cycles as trigger points for opening and closing the gate. As in frequency counting, the measurement is automatically repeated at a rate that is manually variable in commercial instruments, or the measurement may be made on a one-shot basis.

Period measurements allow more accurate measurements of unknown low-frequency signals, because of increased resolution. For example, a frequency measurement of 100 Hz on a counter with 8-digit display and a 10-sec gate time, will be displayed as 0000.1000 kHz. A single period measurement of 100 Hz on the same instrument, with 10 MHz as the counted frequency, would be displayed as 0010000.0 μsec. Thus, resolution is increased by a factor of 100. The accuracy here is also affected by

the ±1-count ambiguity, ± the time-base accuracy, ± the trigger error.

Time-interval measurements are similar to period measurements, with the added capability of setting trigger levels for the gate (Fig. 6-12) to any desired amplitude. Also, two different inputs can be used to start and to stop the counting. Alternatively, the time interval between two setable levels on one waveform can be measured.

Phase measurement is a special time-interval measurement. Usually, it is made by measuring the time between zero crossings of two voltages of the same frequency. Accuracy is usually limited by the inability to trigger the gates precisely at zero crossings.

Multiple-period Averaging. The effects of the ±1-count ambiguity and the trigger error can each be reduced in decade steps by using multiple-period averaging (Fig. 6-13). In one leading high-frequency counter, for example, the function selector switch is ganged to the decade-divider assemblies so that the input signal can be scaled in decade steps by factors up to 100,000 to reduce trigger error. The ±1-count ambiguity is also reduced by a factor of 10 for each decade of scaling selected for the input signal. In the low-frequency measurement example above, the counter would display 10000.000 μsec for a 100-period average. The function selector switch automatically shifts the decimal point in the display to show the correct reading for a single period.

Ratio Measurements. The ratio of two frequencies is determined by using the lower-frequency signal for gate control while the higher-frequency signal is counted, as shown in Fig. 6-14. With proper transducers, ratio measurements may be applied to any phenomena that may be represented by pulses or sine waves. Gear ratios and clutch slippage, as

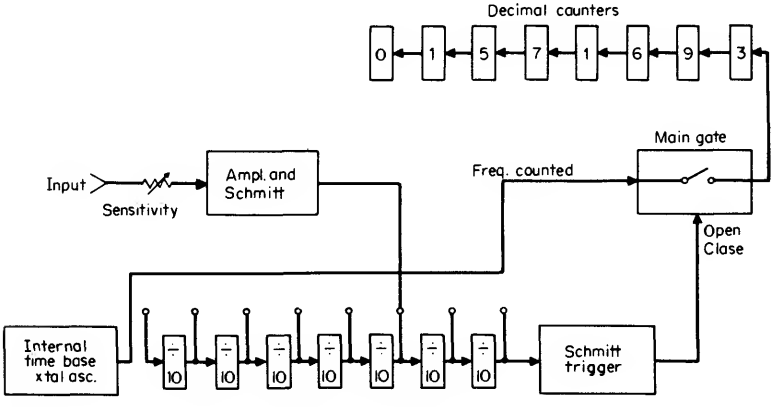


FIG 6-13 Arrangement for multiple period averaging

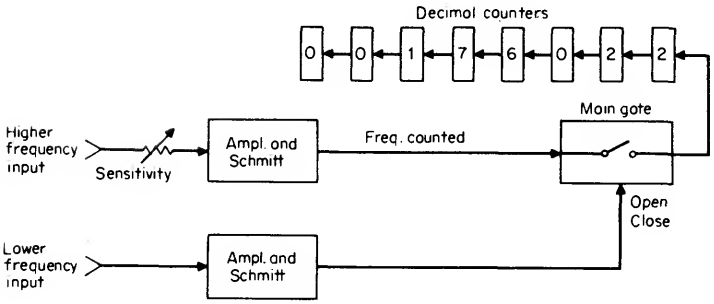


FIG 6-14 Function switch set to measure the ratio of two frequencies.

well as frequency divider or multiplier operation, are some of the measurements which can be made by using this technique.

Accuracy is ± 1 count $\pm$ trigger error. The accuracy may be improved by using the multiple-period averaging technique by counting the higher frequency for 10^n cycles of the lower frequency.

Rate Measurements. With a preset counter or a counter with a preset plug-in, frequency measurements can be normalized automatically to rate measurements by appropriate selection of the gate time. The counter will then display a readout corresponding to the desired engineering units. For example, the Hewlett-Packard 5214L preset counter can be set to a gate time of 600 msec to cause an input from a 100 pulse per revolution tachometer to be displayed directly in revolutions per minute.

High-frequency Measurements. Accurate high-frequency measurements can be made above the normal range of an electronic counter by using heterodyne converters, transfer oscillators, or automatic dividers, and for frequencies up to 500 MHz, prescaling is available. The unique capabilities of each will now be briefly described.

Heterodyne converters enable a counter to measure the average values of continuous wave (CW) signals (even when FM'd to a certain extent) and have a resolution of about 1 Hz/1 sec of counter gate time. For a good example of such a converter, refer to Fig. 6-15. The tuning control selects the 200-MHz harmonic that gives a beat-frequency output which, after prescaling by a factor of 4, is within the 50-MHz counting capability of the counter. At the same time, the counter gate time is extended by a factor of 4 so that direct readout is achieved. The frequency reading on the counter is then added to the setting on the tuning dial to give the unknown frequency.

Transfer oscillators, on the other hand, are more versatile. They can measure FM or pulsed signals, as well as CW signals, over a very wide frequency range and can produce N-hertz resolution in 1-sec counter gate

time, where N is the harmonic number, but require calculations (and perhaps two measurements) and thus need more operator skill and time. Note that accuracy may be less when measuring the carrier frequency of pulsed signals.

In operation, the transfer oscillator generates a variable frequency, which is adjusted so a harmonic of that frequency's zero beats with the unknown CW signal (see Fig. 6-16). The transfer oscillator frequency is then measured on the counter and multiplied by the appropriate harmonic number to give the unknown frequency. In the Hewlett-Packard 2590B, zero beat is obtained by an automatic phase-lock loop after one of the nearest subharmonics has been manually tuned. Measurements to 15 GHz are possible with the 2590B model, and to 40 GHz with the Hewlett-Packard 540B with related instruments. The figures are somewhat misleading unless it is realized that the mixer, harmonic generator, and oscillator are common components for both instruments.

Automatic frequency dividers provide automatic measurement and direct readout of a wide range of CW frequencies, and typically furnish 1,000-Hz resolution in 1 sec. Prescaling is accomplished by frequency division of the input signal. If the gate time is extended with the scale factor, the correct frequency will appear on the counter readout. Because the prescaler is a wideband instrument, it is more susceptible to noise than

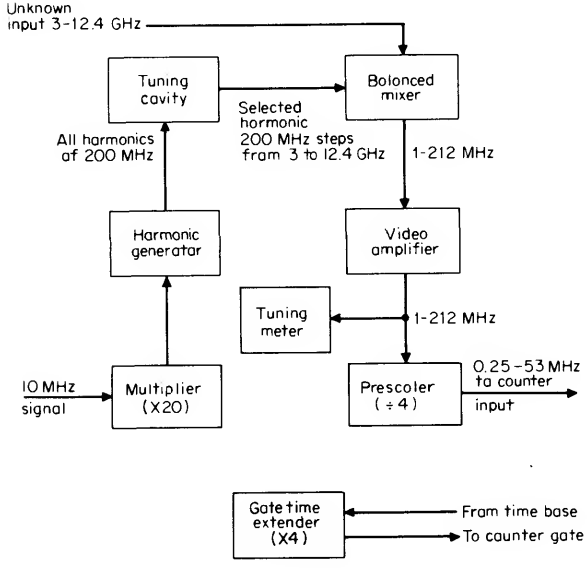


FIG 6-15 Heterodyne scheme to count higher frequencies.

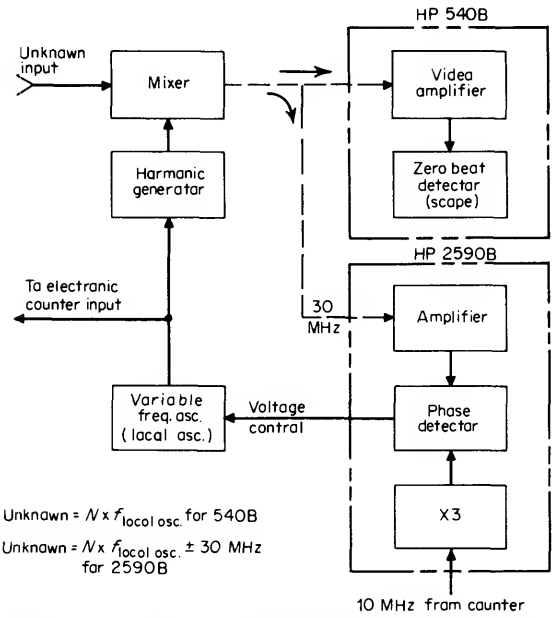


FIG 6-16 Use of a transfer oscillator to measure higher frequencies. (Courtesy of the Hewlett-Packard Co.)

tuned instruments like the heterodyne converters. An adjustable trigger level control on the prescaler can be used to discriminate against unwanted signals. The accuracy of the prescaler is the same as that of the counter, although the measurement takes 2, 4, or 8 times as long in time, depending upon the scale factor.

6-5 *Frequency Synthesizers*

A great deal of the information in this section was excerpted, with permission from the publisher, from a paper by V. E. Van Duzer [15]. It could be argued that a frequency synthesizer should be classified as a signal source rather than as a time-and-frequency instrument, but the *emphasis* in synthesizers is on precision and stability as much as on variability. The instrument, in the best designs, is essentially a variable *frequency standard.*

Sources having variable frequency, precisely setable and with stabilities comparable with those of good frequency standards, are valuable in highly developed communications work, radio distance sounding, radar, doppler systems, automatic and manual testing of frequency-sensitive devices,

numerous timing situations, spectrum analysis, stability studies, and many other areas.

Frequency standards have already been discussed. Their availability and excellent performance stimulated a search for ways to translate their stability to any desired frequency. This translation, when the operation is something more than a single fixed operation, is commonly known as *frequency synthesis*. Hence, a variable-frequency synthesizer is an instrument that translates the frequency stability of a single frequency, usually one from a frequency standard, to any one of many other possible frequencies, usually over a broad spectrum. Such an instrument may provide any one of thousands, even billions, of frequencies. In everyday usage the word *variable* is usually omitted from the name, and the instruments are merely called *frequency synthesizers*.

The two basic approaches to frequency synthesis are known as *direct* (or *true*) and *indirect*. Direct synthesis simply performs a series of arithmetic operations on the signal from the frequency standard to achieve the desired output frequency. The indirect method uses tunable oscillators, which are phase-locked to harmonics of signals derived from the standard [16].

The direct-synthesis approach has the pronounced advantages of permitting fine resolution and fast switching in the same instrument and fail-safe operation and an extremely clean output signal as well. This is the type of instrument to be discussed in this section.

Basically this system started with a stable oscillator or frequency standard that provided a signal of frequency f_0, say 1 MHz. Frequency multipliers and dividers were used to give such frequencies as 0.1 MHz and all integral multiples of 1 MHz from 1 to 10. The 0.1-MHz signal and the 10-MHz signal were also multiplied by integers from 1 to 10. Now, if 23.6 MHz were desired, the 20-, the 3-, and the 0.6-MHz signals could be applied to suitable modulators to give the sideband sum of 23.6 MHz.

The above approach sounds very simple, but none of the signals mentioned is pure, and besides, modulators (or mixers) always produce unwanted sidebands or other spurious frequencies. Therefore, the separation of the desired frequency from the unwanted ones is a serious problem, especially when one wants to select one of many output frequencies spaced closely together. As a rule of thumb, it has been considered impractical to filter out spurious components that fall within 10 percent of the desired output frequency. This difficulty has stimulated many searches for better methods of multiplication, division, mixing, and filtering. A good summary of early work is given in Technical Report 2271 of the United States Army Electronics Research and Development Laboratory, Fort Monmouth, New Jersey. If still in print, it can be

obtained from the Office of Technical Services, U.S. Department of Commerce, Washington, D.C.

For tutorial purposes, it is thought best to describe a modern synthesizer with the capability of providing an extremely large number of frequencies and high output quality.

The design objective is to obtain a very large number of switchable frequencies from a single stable signal, frequencies arranged in decimal fashion, as on the keyboard of an adding machine. One synthesizer, for example, provides frequencies from 0.01 Hz to 50 MHz in digital increments as fine as 0.01 Hz—a total of 5 billion discrete frequencies. At any frequency the output is a spectrally pure signal. Any nonharmonic spurious signal is more than 90 dB below the desired signal (see Fig. 6-17). The output frequency can be selected by front-panel pushbuttons or by remote electronic control. When the frequencies are electrically programmed, switching can be accomplished in less than 1 msec.

A simplified block diagram of the overall instrument is shown in Fig. 6-18. The driver contains a frequency standard, a spectrum generator, and appropriate selection networks to provide a series of fixed frequencies

FIG 6-17 Frequency synthesizer with output from 0.01 Hz to 50 MHz in 0.01 Hz increments. (Hewlett-Packard Model 5100A)

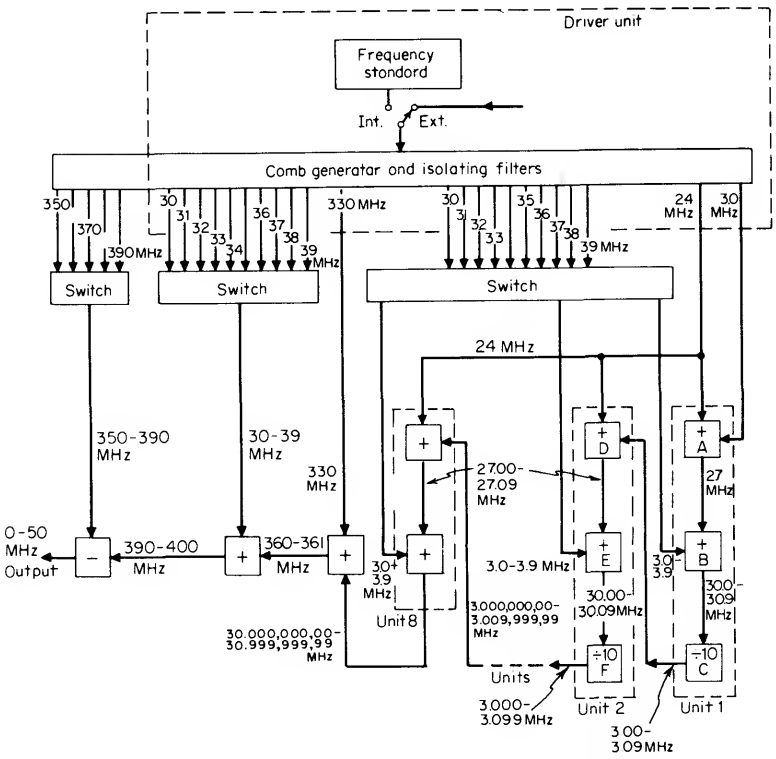


FIG 6-18 Basic circuit arrangement for the frequency synthesizer shown in Fig. 6-17.

between 3 and 39 MHz to the synthesizer unit. The synthesizer unit contains harmonic generators and suitable mixers, dividers, filters, and amplifiers to derive the desired output frequency as a function of the fixed frequencies.

The fine-resolution portion of the instrument is particularly interesting and also serves to illustrate the method of synthesis used. As shown in the right-hand portion of Fig. 6-18, there are seven identical mixer-divider units, each of which corresponds to a place, or decimal position, in the final output frequency number. In each of these units, and in the eighth unit as well, a frequency of 24 MHz is used as a carrier input, as shown.

In the right-hand unit, which produces what ultimately becomes the highest resolution digit (10^{-2} Hz), the 24-MHz carrier is added to a 3.0-MHz frequency in frequency adder *A* to produce 27.0 MHz. In *B*

the 27.0-MHz frequency is added to a frequency of from 3.0 to 3.9 MHz, depending on the setting of the panel pushbutton or remote-control circuit. Selection of a 2 in this particular digit position, for example, electronically selects a signal of 3.2 MHz from the driver.

The output of *B* is a frequency of 30.0 to 30.9 MHz, which is divided in *C* to produce 3.00 to 3.09 MHz. This frequency is applied to the second unit, where it adds with the 24-MHz carrier as before, and the process repeats. If the process is followed through, it will be seen that the frequencies noted in the block diagram are obtained at the outputs of the various adders and dividers. In essence, each mixer-divider unit, through a frequency-division process, moves a given digit one place to the right for the final frequency of between 30,000,000.00 and 30,999,999.99 Hz, depending on the output frequency selected.

In the following two operations the signal is added to a frequency of 330 MHz, and the resultant is again added to an appropriate frequency between 30 and 39 MHz to yield a frequency of between 390 and 400 MHz. One of the five frequencies from 350 to 390 MHz is then subtracted from this to yield the desired 0.01-Hz to 50-MHz output frequency. A good way to start understanding the operation of the circuit is to observe that a frequency change of 0.1 MHz in the input to unit 1 produces a change of 0.01 Hz at the output of the whole system, because of the repeated division by 10.

The design of the synthesizer described resulted from a long effort to optimize performance for a reasonable cost. The design is therefore arbitrary to some extent, and other choices of frequencies and mixing methods are possible.

Applications of Synthesizers. If in the digital frequency synthesizer we have a frequency standard whose output frequency can be selected by either manual or electronic command to very high resolution in less than a millisecond, such an instrument constitutes a most powerful tool. In communications work, for example, the synthesizer's excellent spurious-frequency performance makes it well suited to use as the master oscillator in a transmitter and as the local oscillator in a receiver. If the transmitter and the rf section of the receiver are untuned, an extremely fast switching system can be used to change the local oscillator (synthesizer) frequency to achieve communications systems of high performance.

Again, the synthesizer can greatly facilitate surveillance work if it is used as the local oscillator in a receiver designed to determine accurately the frequencies of remote transmitters. The ease and speed with which the synthesizer frequency can be changed allow monitoring of a multiplicity of channels with a single receiver by sequencing the local oscillator (synthesizer) through the desired channels.

Sequencing the synthesizer output through a group of desired fre-

quencies also permits a single instrument to operate as an automatic calibrator for a multiple-frequency setup such as a multiple-transmitter installation. The arrangement can provide for phase-locking the transmitter frequencies to the synthesizer by a circuit with a time constant long enough to maintain the transmitter frequency for the duration of the sequencing cycle.

The effective use of the microwave spectrum for communications requires frequency sources with extremely good fractional frequency stability so that the receiver bandwidth can be minimized. With a 3-kHz information bandwidth at 10 GHz, a frequency stability of 3 parts in 10^8 for a duration of a message is desirable for double-sideband work. For single-sideband work the requirement is about 1 part in 10^9 for the same conditions. Obviously, a synthesizer must be used in such a communications system to make it practical.

Determining the velocity of distant space vehicles through doppler frequency measurements involves operation at x-band with receivers having intermediate-frequency bandwidths of but a few cycles to minimize noise levels. As the vehicle velocity changes, the receiver's local oscillator frequency must be changed to keep the received signal in the center of the intermediate-frequency bandwidth. Here again, the synthesizer is ideal because of its stability and because its frequency can be changed in known and selectable increments.

Finally, the synthesizer is indispensable for automatic testing schemes in which signals having specific frequencies must be rapidly programmed into a test setup.

CITED REFERENCES

1. U.S. National Bureau of Standards: *Proc. IRE*, vol. 48, pp. 105–106, January, 1960.
2. Hartke, D.: A VLF Comparator for Relating Local Frequency to U.S. Standards, *Hewlett-Packard J.*, vol. 16. no. 2, October, 1964.
3. S. Flügge-Freiburg, ed.: *Encyclopedia of Physics*, vol. 23, pp. 289–372, Springer-Verlag OHG, Berlin, 1966.
4. Frequency and Time Standards, Hewlett-Packard Co., Application Note 52, Palo Alto, Calif.
5. Cutler, L. S.: Some Aspects of the Theory and Measurement of Frequency Fluctuations in Frequency Standards, in *Proc. Symp. Definition and Measurement of Short-term Frequency Stability, Goddard Space Flight Center, Greenbelt, Md., Nov. 23–24, 1964*, under Auspices of U.S. NASA and IEEE, NASA SP-80, pp. 89–100,
6. Davenport, W. B., and W. L. Root: "Introduction to Random Signals and Noise," secs. 4 and 5, McGraw-Hill Book Company, New York, 1958.
7. *Ibid.*, chap. 6.
8. Cady, Walter G.: "Piezoelectricity; an Introduction to the Theory and Applications of Electromechanical Phenomena in Crystals," Dover Publications, Inc., New York, 1964.

9. Bottom, Virgil E.: "The Theory and Design of Quartz Crystal Units; an Introduction to the Basic Principles of Piezoelectricity and Their Application to the Design of Quartz Crystal Units," MacMurray Press, Jacksonville, Ill., 1968.
10. Hammond, D. C., C. Adams, and L. S. Cutler: Precision Crystal Units, *Frequency*, p. 29, July–August, 1963.
11. S. Flügge-Freiburg, *op. cit.*, p. 321.
12. Davidovits, P.: An Optically Pumped Rb^{87} Maser Oscillator, *Proc. Symp. Definition and Measurement of Short-term Frequency Stability, Goddard Space Flight Center, Greenbelt, Md., Nov. 23–24, 1964*, NASA SP-80, p. 171, 1965.
13. Barnes, J. A., D. W. Allan, and A. E. Wainwright: *IRE Trans. Instr., I-II*, vol. 26, 1962.
14. For example, Operating and Service Manual—Electronic Counter 5233L, Hewlett-Packard Co., Palo Alto, Calif.
15. Van Duzer, Victor E.: A 0–50 Mc Frequency Synthesizer with Excellent Stability, Fast Switching, and Time Resolution, *Hewlett-Packard J.*, vol. 15, no. 9, May, 1964.
16. Hughes, R. J., and R. J. Sacha: The LOHAP Frequency Synthesizer, *Frequency*, vol. 6, no. 8, pp. 12–21, August, 1968.
17. Cutler, L. S.: Some Aspects of the Theory and Measurement of Frequency Fluctuations in Frequency Standards, *Proc. IEEE*, vol. 54, no. 2, February, 1966.

CHAPTER SEVEN

DIRECT-CURRENT INSTRUMENT AMPLIFIERS

From Notes by

Paul Baird, Bill Kay, Craig Walter, and Richard Y. Moss, II

Hewlett-Packard Company, Loveland, Colorado

Amplifiers of many kinds are used in electronic instruments. Also, instruments are used in many ways to make measurements *on* amplifiers. It is not feasible to present a full treatise on amplifier design in this volume, but some of the dc amplifiers in instruments have such special requirements that this chapter will be devoted to them.

The principal methods for making measurements on amplifiers by means of electronic instruments will be presented in Chap. 13, along with a short discussion of the characteristics that need to be measured.

Whether they are in instruments or not, amplifiers are employed for one or more of the following reasons: (1) to increase the power available in an electrical signal, (2) to amplify voltage or current levels where power per se is not of great concern, (3) automatically to *limit* the voltage or current that can be delivered to a load, (4) to provide a prescribed

transfer function, either linear or nonlinear, between a source and a load, (5) to provide the desired load impedance on a source or the desired source impedance for a load, and (6) to attenuate or reject the common-mode component of voltage on a pair of conductors (the common-mode voltage is the *average* of the voltages on the two conductors at each instant with respect to ground or some designated reference potential). Amplifiers are frequently used as active filters, but this is a special use under item 4.

In instruments, dc amplifiers are most commonly used for reasons 2 and 5 above. When a signal is very small, as a voltage, as a current, or as both, one needs amplification. Whether one needs a high voltage gain E_L/E_s, a high current gain I_L/I_s, a high transconductance I_L/E_s, or a high transimpedance E_L/I_s depends upon the natures of the signal source and load to be driven. The simple Thevenin and Norton equivalent circuits in Fig. 7-1, using an amplifier with a common ground terminal, will help to clarify this situation. The symbols are defined on the figure. If a signal is most accurately described by its Thevenin equivalent voltage E_s even when Z_s is variable, then Z_i of the amplifier should be high compared with Z_s. If the signal is most accurately described as a current I_s, regardless of variations in Z_s, then one desires $Z_i \ll Z_s$. Similar considerations determine the desirability of $Z_o \gg Z_L$ or $Z_L \gg Z_o$.

So far, the situation seems very simple, for shunt or series feedback can be used to modify Z_i and Z_o over wide limits while stabilizing the desired transfer gain. The common problem, however, is that dc signal sources can be at such low levels that amplifier noise and drift in operating points degrade the accuracy of measurement. This chapter describes some of the methods used to deal with drift and noise while simultaneously achieving the desired gains, bandwidths, and impedance relationships. The role of bandwidth in the design of dc instrument amplifiers is important; the bandwidth is of course never zero, since a signal must vary to contain intelligence.

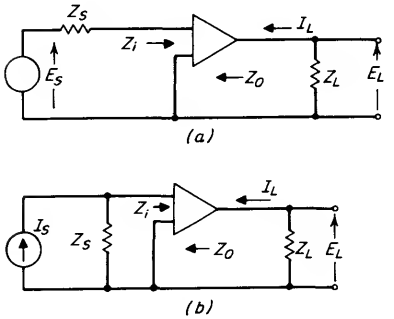


FIG 7-1 Thevenin (a) and Norton (b) equivalent amplifier circuits.

The most straightforward way to make the frequency response of an amplifier extend down to 0 Hz is to couple the stages conductively. The resulting amplifier is said to be *direct coupled* and besides being straightforward in concept, it can have essentially the same upper frequency range as one with capacitor coupling. An example of a direct-coupled amplifier will be given in Sec. 7-1 to show some of the design considerations.

However, it is difficult and expensive to keep the quiescent operating potentials in the direct-coupled amplifier from drifting with age and temperature. Also, many input stages, such as electron tubes and transistors, exhibit excess random noise as the frequency approaches zero. In some applications it is appropriate to use either periodic resetting of the operating point or modulation (chopping) of the input signal. Chopping converts the frequency band of the input information to some ac carrier band and avoids the use of direct coupling. Resetting and chopping will be presented later in the chapter.

7-1 Direct-coupled Amplifier Considerations

The most common means of amplifying dc voltages, excluding amplification of extremely low level signals of the order of 1 μV, is the direct-coupled amplifier. The amplifier can be either inverting (operational) or noninverting.

If, in addition to the requirement of accurate low-level amplification, the amplifier is required to have moderate bandwidth (20 kHz at a gain of 40 dB), wide dynamic range (0 to $\pm$15V) at the input, extremely high input resistance (10^{10} Ω), very low offset voltage and current (<1 μV and 1 pA at the amplifier input) and remain relatively insensitive to its environment (power supplies, source and load impedances, temperature, and humidity), then its design is greatly complicated.

Chopper stabilization as a technique does not normally satisfy the bandwidth requirement. Such amplifiers are normally relegated to amplification of very low frequency voltages near 0 Hz. To amplify the higher frequencies, an ac-coupled amplifier can be paralleled, but this is of course more complicated and more costly.

Up-conversion to a much higher frequency carrier (megahertz) would accomplish amplification of the frequency range required, but amplifier accuracy and dc stability (variation of the offset voltage and current at the amplifier input with time and temperature) are difficult to achieve with available chopping devices at high frequencies.

Amplification by means of a direct-coupled amplifier, then, remains the simplest and least-expensive means for satisfying the requirements previ-

ously listed. The following few pages detail a case history of the design and evaluation of an amplifier with common ground used to provide signal processing from a measurement transducer to a moderate-speed (15-kHz bit rate) analog-to-digital converter. The analog-to-digital converter in this case was actually a digital voltmeter.

Basic Configuration. To achieve accurate and stable gain, and high input

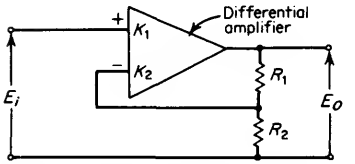


FIG 7-2 Basic configuration of the direct-coupled amplifier.

impedance, a noninverting configuration with voltage feedback was chosen. The amplifier shown in Fig. 7-2 is basically a good *differential* amplifier with the noninverting gain K_1 made as close as possible to the inverting gain K_2. The desirability of this arrangement will become clear.

As already noted, one of the chief design problems is to reduce sensitivity to environment. Such a reduction in sensitivity to environment will not be achieved, however, unless the fractional changes in both forward gains, as a function of input level, time, or temperature, are identical. That is to say, the differences in the gains must be small or the common mode rejection (CMR) ratio must be low to obtain ultimate benefit from the feedback. The definition is

$$\text{CMR} = \frac{\text{voltage gain, common inputs}}{\text{voltage gain, differential inputs}} = \frac{K_1 - K_2}{K_1} \approx \frac{K_1 - K_2}{K_2}$$

Input Stages. A matched pair of field-effect transistors (FETs) was chosen as the differential input stage. The FET offered both the high input resistance and the small leakage current required. Bipolar devices, though not necessarily limited by their lower input resistance since their input resistance is boosted by the amplifier feedback, have considerably more leakage current—the transistor base current is typically 100 to 1,000 times more than the FET gate leakage current.

The FETs are operated in a balanced common-drain configuration to achieve minimum sensitivities of the input offset voltage to power supply,

device parameter, and temperature variation (the input offset ΔV_{GS} is the required difference in voltages at the input terminals required to give zero differential output). Figure 7-3 is a simplified diagram of the circuit arrangement. To further reduce offset variations caused by temperature fluctuations, the FET environment is temperature controlled at a temperature higher than the maximum expected ambient. Although this does increase the gate leakage current, it also keeps it constant. Minimum sensitivity of gain to device parameter variation is also achieved.

The common drain must be made to track or bootstrap the input voltages to maintain V_{DS} constant with input level. If this is not done, the offset voltage ΔV_{GS} will change as a function of level because of the changing FET bias. The back-biased breakdown (Zener) diode in Fig. 7-3 acts as if a battery were connected between the emitters of transistors Q2A and Q2B and the two drain terminals; and since the base-to-emitter voltages are very small, V_{DS} for both FETs is kept nearly constant, even though a large common mode input may exist.

Current sources for the active devices are used to reduce the effects of varying power-supply voltages and bias the FETs in such a way as to keep V_{DS} and I_{DS} fixed while V_{GS} varies. Actually, the current sources in the amplifier being described are active circuits. A very high value of resistance could be used to create a constant current source, but this approach would require higher power-supply voltages.

Although the FETs used had much less gate leakage current than the base current of a transistor, it was still much larger than desired. A

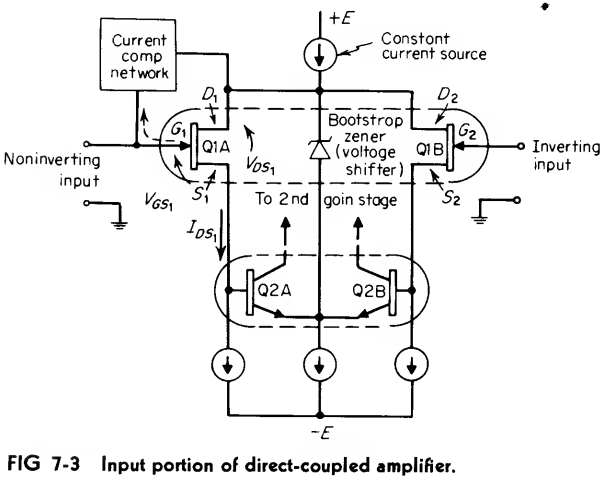


FIG 7-3 Input portion of direct-coupled amplifier.

constant current source bootstrapped to the drain potential achieved good compensation without lowering the incremental input impedance.

Because the amplifier does dissipate power, albeit small, temperature gradients exist across and through it. The temperature differences, although typically only a few degrees, may cause severe problems because of thermoelectric effects in various parts of the amplifier. Typically, transistors that are hermetically packaged are sealed in an envelope made of Kovar. The leads are also Kovar—penetrating the envelope header through glass. Kovar is used because its coefficient of thermal expansion is nearly equal to that of glass, a match that is made necessary by the elevated temperatures used to bond glass to Kovar and thereby to obtain hermeticity. A thermocouple junction is formed wherever the Kovar lead connects to a dissimilar metal. Since copper is used almost exclusively for the conductors on printed circuit boards, Kovar-copper thermocouples exist wherever transistors are soldered to the board. If these junctions are at different temperatures (caused by the aforementioned gradients), different voltages will exist across them. These voltages will, of course, change whenever the temperatures of the junctions change.

If one assumes the voltage generated between two of these couples, separated by 1°C, to be 150 μV, then temperature fluctuations as small as 0.01°C can create measurable and very troublesome errors. Great care must be taken with the physical layout of the input stages to ensure either the exclusion or the stability of these temperature gradients. The amplifier input stages must be so isolated that surrounding air is not allowed free movement.

These effects are particularly troublesome when an attempt is made to evaluate the actual temperature coefficient of the offset voltage. The amplifier is placed in an oven and its temperature varied. Adequate time must be allowed for the stabilization of temperature gradients. Even in a small environmental chamber this time can be hours. The voltage must be continuously monitored to ensure stabilization.

The amplifier environment must also be rigidly controlled when stability and noise measurements are being made, so that variations in offset are truly caused by aging rather than by temperature.

In Fig. 7-2 the gain of the amplifier with feedback is

$$K' = E_o/E_i = \frac{K_1}{1 + K_2\beta} \approx \frac{1}{\beta}\Bigg]_{\substack{K_2\beta \gg 1 \\ K_1 = K_2}} \tag{7-1-1}$$

where K_1 = noninverting gain
K_2 = inverting gain
$\beta = R_2/(R_1 + R_2)$

The gain is set accurately by adjusting β.

If aging or some environmental change causes K_1 and K_2 to change, the relative change in K' is

$$\frac{\Delta K'}{K'} = \frac{(1 + K_2\beta)\,\Delta K_1 - K_1\beta\,\Delta K_2}{(1 + K_2\beta)^2}\,\frac{1 + K_2\beta}{K_1}$$

$$= \frac{1}{K_1(1 + K_2\beta)}\,(\Delta K_1 + K_2\beta\,\Delta K_1 - K_1\beta\,\Delta K_2) \qquad (7\text{-}1\text{-}2)$$

If $K_2\beta \gg 1$,

$$\frac{\Delta K'}{K'} \equiv \frac{\Delta K_1/K_1}{K_2\beta} + \frac{\Delta K_1}{K_1} - \frac{\Delta K_2}{K_2} \qquad (7\text{-}1\text{-}3)$$

and if $\Delta K_1/K_1 = \Delta K_2/K_2$,

$$\frac{\Delta K'}{K'} = \frac{\Delta K}{K}\,\frac{1}{K\beta} \qquad (7\text{-}1\text{-}4)$$

where $K = K_1 = K_2$.

Gain Stages. Two differential transistor amplifier stages provide the forward gain of the whole amplifier, as shown in Fig. 7-4. The common-drain FETs in Fig. 7-3 act only as differential impedance transformers

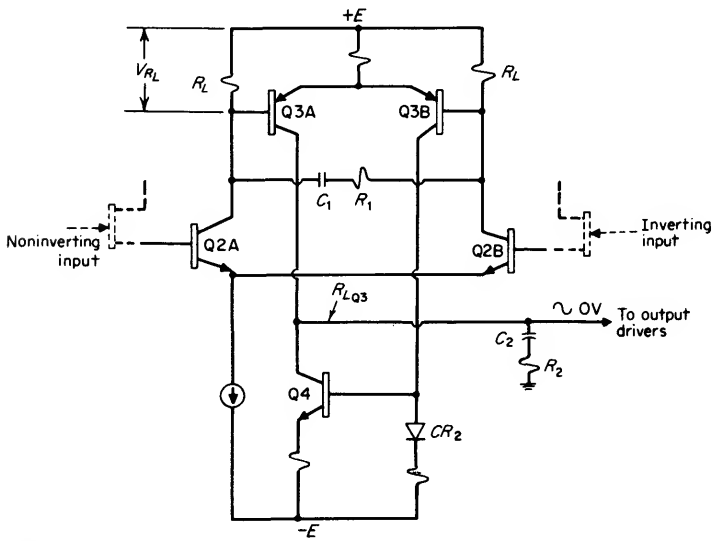


FIG 7-4 Gain stages.

with voltage gain slightly less than unity. Observe that Q2A and Q2B in Fig. 7-4 are the transistors shown in Fig. 7-3; they are repeated for clarity.

Both n-p-n and p-n-p transistor pairs are used in complementary fashion to make the operating voltage at the output approximately ground potential. These pairs must be well matched to produce low drift and good CMR. The pair Q2A and Q2B are especially critical since they operate at input signal level. The Q3 pair is less critical in regard to matching. For detailed analysis of transistor amplifiers, see recent specialized books and papers [1].

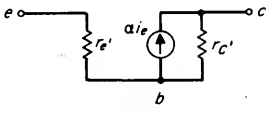


FIG 7-5 Simplified common emitter circuit.

Noise considerations (and the limits on power-supply voltages) indirectly limit the gain of Q2. The two major contributors to amplifier noise (excluding the $1/f$ noise of the FETs) are the equivalent-noise voltage of Q2 and the equivalent-noise current of Q3. This noise voltage E_N, referred to the input and assuming no correlation, is given by

$$E_N^2 = E_{NQ2}^2 + \left(\frac{I_{NQ3}R_L}{K_{Q2}}\right)^2 = E_{N^2Q2} + (I_{NQ3}r'_{eQ2})^2 \tag{7-1-5}$$

with the assumption $R_L \ll r'_{eQ2}$ (Fig. 7-4), and where r'_e and r'_c are the equivalent resistances in the simplified common-base circuit of Fig. 7-5. The forward voltage gain of the amplifier is

$$K \approx \frac{R_{LQ3}}{2r'_{eQ3}} \frac{R_L}{r'_{eQ2}} \tag{7-1-6}$$

where $R_L \ll 2\beta_{Q2}r'_{eQ3}$ and $\beta = \alpha/(\alpha - 1)$.

Since r'_{eQ2} is almost inversely proportional to the emitter-bias current I_{EQ2}, an increase in this operating current will apparently increase the gain and decrease the total noise. This improvement is limited, however, by power-supply voltage, since for a given R_L, the voltage drop across it is proportional to the current in Q2. Once R_L and r'_{eQ2} are determined by the practical limitations, then

$$V_{R_L} \approx I_{EQ2}R_L = \frac{CR_L}{r'_{eQ2}} \tag{7-1-7}$$

and

$$r'_{eQ2} = \frac{CR_L}{V_{R_L}} = C'R_L \tag{7-1-8}$$

where $C = I_{EQ2}r_{eQ2}$.

Therefore,

$$E_N^2 = E_{NQ2}^2 + (I_{NQ3}C'R_L)^2 \tag{7-1-9}$$

and

$$K = \frac{R_{LQ3}}{2C'r'_{eQ3}} \tag{7-1-10}$$

The resistance R_L cannot be made vanishingly small, however, for as R_L decreases, r'_{eQ2} must decrease (and I_{EQ2} must increase) to maintain K. Unfortunately, E_{NQ2}^2 is a very strong function of I_{EQ2}, increasing with increasing I_{EQ2}. For fixed supply voltages, then, there is an optimum value for R_L.

The RC pairs (R_1 and C_1, R_2 and C_2) shown in Fig. 7-4 are used to shape the amplifier frequency response. The size of capacitor used is primarily limited by the slewing rate required. The slew limit is the maximum rate of change of signal $de/dt_{\max}$ which the amplifier can follow linearly. Slew limiting occurs when the current required to produce a voltage across a capacitor, $i = C(de/dt)$, exceeds the maximum current available. If the current has been fixed by gain and noise considerations and the signal voltage and frequency required are known, the maximum capacitor size is determined. This presents a real restraint on the size of C_2 because of the large voltage swings required (± 15 V). Therefore C_2 was determined by slew rate considerations; R_1, R_2, and C_1 were then chosen to achieve the desired frequency response.

The transition from a balanced or differential configuration to a single-ended one is effected through Q4. Such a stage is necessary to achieve both the high gain and the power-supply rejection required. Diode CR_2 provides a first-order correction to change in the base-emitter voltage of Q4, caused by temperature.

Output Stages. Cascaded complementary emitter followers are used to couple from the collectors of Q3A and Q4 to the output. This configuration provides both the high input impedance required for the output stage (>10 MΩ) and relatively good linearity near zero voltage. The resistors in the collectors and emitters of the output transistors limit the output current and thereby provide short-circuit protection. The design of this stage is fairly straightforward and will not be discussed here. See Fig. 7-6 for the simplified schematic.

Overvoltage Protection. In any measuring environment in which measurements are not rigorously controlled and monitored, the possibility exists of damaging the amplifier through input overvoltage. To prevent damage to the amplifier (and perhaps to the signal source), dynamic protection must be provided. See Fig. 7-7.

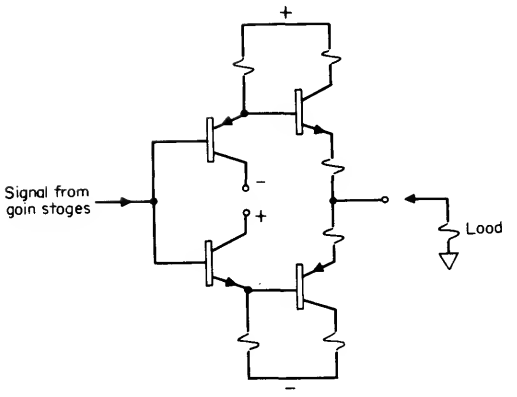


FIG 7-6 Output stages.

When the dynamic range of the amplifier is exceeded, the diodes will conduct. Current limiting is provided by the resistor in series with the input. The top pair of diodes do not significantly degrade the input impedance because of bootstrapping by the loop gain. The primary disadvantage is an increase in noise (the thermal noise of the resistor).

Open-loop Gain Measurement. Open-loop gain measurements on dc amplifiers are usually exceedingly tedious and difficult because of the absolute magnitude of K (>100 dB). The output swing is limited and may be exceeded because of voltages generated by noise, thermals, or component temperature sensitivity at the amplifier input. (If the amplifier voltage gain is 120 dB, then 1 μV at the input will generate 1 V at the output. If the amplifier is relatively broadband, its noise, referred to the input—tens of microvolts, peak to peak—may be large enough to saturate the amplifier output.) Such open-loop gain measurements are necessary if accurate determination of the frequency response and the shaping needed to ensure stability are to be made.

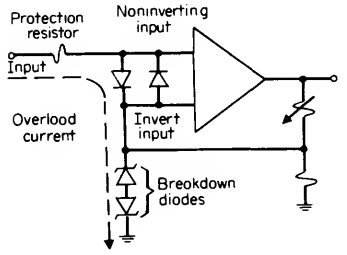


FIG 7-7 Protection against excessive input voltages.

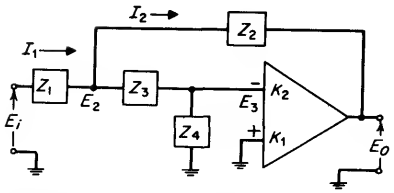


FIG 7-8 A test arrangement for measuring forward gain.

Within the assumptions noted, the method shown above (Fig. 7-8) is both accurate and relatively effortless.

Let $K_1 = K_2 = K$. The transfer from the amplifier input to its output is given by

$$E_o = -KE_3 = -K\frac{Z_4}{Z_3 + Z_4}E_2 \tag{7-1-11}$$

Because of the feedback current through Z_2, then $E_2 \ll E_i$, and

$$\frac{E_i - E_2}{Z_1} \approx \frac{E_2 - E_o}{Z_2} \tag{7-1-12}$$

$$E_i = E_2\left(\frac{Z_1}{Z_2} + 1\right) - \frac{Z_1}{Z_2}E_o \tag{7-1-13}$$

Adding Eqs. (7-1-11) and (7-1-13),

$$\frac{E_i}{E_2} = \frac{Z_1}{Z_2} + 1 + K\frac{Z_1Z_4}{Z_2(Z_3 + Z_4)} \tag{7-1-14}$$

Let $Z_1 = Z_2$, $(Z_3 + Z_4)/Z_4 = 100$, and $K = 200$. Then

$$\frac{E_i}{E_2} \approx \frac{K}{100} \tag{7-1-15}$$

and

$$K = 100\frac{E_i}{E_2} \tag{7-1-16}$$

Output Sensing. A common measurement error occurs when measuring gain accuracy or linearity if the resistance in a length of wire, of any size, is neglected. Any resistance in series with the output of the amplifier (and with its load) will account for some voltage drop and thereby reduce

the output voltage and introduce an apparent error in amplifier gain. A few milliohms can result in significant errors when extremely accurate (0.001 percent) gain is required.

As shown in Fig. 7-9*a*, the lead resistance R is not included in the feedback loop when remote output sensing is not employed. In *b*, however, the output terminals are made the reference terminals for the voltage feedback and also the ground returns of both the amplifier and the input signal. In this arrangement, R simply becomes part of the output resistance of the amplifier, which is reduced in effect by a factor of $1 + \beta K$ by the feedback.

Results. Typical operating specifications for the amplifier as previously described are given below, and a circuit diagram is shown in Fig. 7-10.

GAIN: $\times 1$ to $\times 100$
GAIN ACCURACY: ± 0.005 percent
LINEARITY: ± 0.001 percent, 0 to ± 15 V
ZERO DRIFT REFERRED TO INPUT:
  Offset voltage ≤ 0.5 μV/°C, 0 to 60°C
  Input current ≤ 1 pA/°C, 0 to 60°C
NOISE REFERRED TO INPUT: 0 to 20 kHz, <8-μV rms
INPUT RESISTANCE: $>10^{10}$ Ω
BANDWIDTH, 3 dB:
  0 to 20 kHz for $\times 100$ gain setting
  0 to 1 MHz for $\times 1$ gain setting
SETTLING TIME: <50 μsec to within 0.01 percent of final value
OUTPUT: ± 15 V max, 0 to 10 mA
POWER SUPPLY REJECTION: ≥ 90 dB for either supply

7-2 *Direct-current Amplifier with Automatic Reset*

Earlier in the chapter it was said that a chopper amplifier could have very low drift, but that a direct-coupled amplifier has advantages when large bandwidth is required. The frequency response, or transfer function, of a direct-coupled amplifier is easily adjusted for best stability and overall response when feedback is used. In contrast, a chopper amplifier,

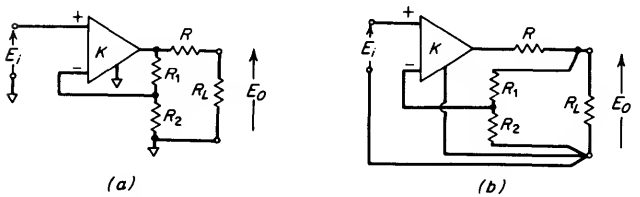


FIG 7-9 Illustration of remote output sensing.

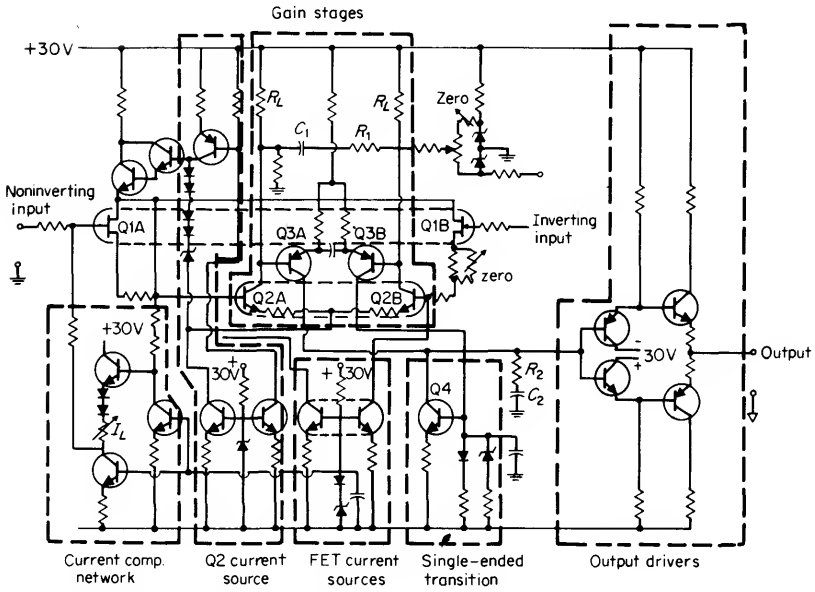


FIG 7-10 Complete amplifier diagram.

Sec. 7-3, has a narrow bandwidth and a rather complicated transfer function. It can be paralleled by an ac amplifier to extend the bandwidth, but when feedback is then used to obtain a flat frequency response, some severe problems appear in stability, transient and overload response, and intermodulation with the chopping frequency.

To generalize, the advantages of the chopper amplifier are (1) that the noise of the amplifier is basically determined by the noise of the active input devices at the chopping frequency rather than at dc, (2) that the active devices can be ac coupled to the input circuit, which greatly reduces leakage currents, and (3) that the voltage offsets of the amplifier are determined by the chopper itself plus the offsets generated by thermal junctions in the input circuitry. However, the development of FETs with little low-frequency noise (10 nV/Hz$^{1/2}$ at 10 Hz) and also low leakage current (10 pA) has changed this situation to some extent.

Slow drift is still a problem in direct-coupled amplifiers, but in some applications automatic reset can be used, as described below, to make the direct-coupled amplifier compete with the chopper amplifier in drift stability.

The Basic Circuits for Automatic Reset. United States patent number 2,994,825 describes means for automatically resetting the operating point

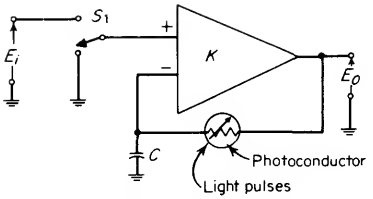


FIG 7-11 Basic automatic-reset amplifier.

of a direct-coupled amplifier periodically on a time-sharing basis. The measurement of the input signal is briefly interrupted every t_0 sec while the amplifier operating conditions are quickly reset to make the output voltage zero when the input is zero. In other words, the drift is balanced out.

In a sense, the amplifier is no longer truly direct coupled, since the signal is briefly interrupted, but in some applications this is no disadvantage.

The original circuit, simplified, is shown in Fig. 7-11. While the input voltage is being amplified, with switch S, in the upper position, the photoconductor is dark and virtually an open circuit. The charge on capacitor C remains practically constant during this interval. Occasionally, when it is permissible to interrupt the amplifying process, the main input terminal of the amplifier is connected to ground by S, and the output is fed back degeneratively through the illuminated photoconductor to C, which is connected to the other input terminal of the differential amplifier.

During reset, the conditions shown in Fig. 7-12 are approached. The offset voltage E_{os} (the voltage that would be required across the input terminals to bring the output to zero) is represented by an equivalent battery, and the photoconductor is illuminated to have a resistance R of about 1,000 Ω. If an ideal amplifier and no initial charge on C are

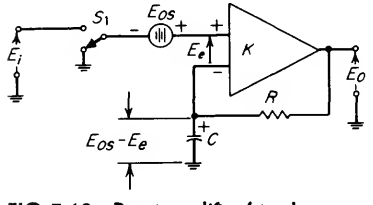


FIG 7-12 Reset amplifier (steady-state conditions) during the reset interval.

assumed, the voltage across C as a function of s, the Laplace operator, is

$$E_C(s) = \frac{E_{OS}}{(1 + K)/K + sRC/K} \tag{7-2-1}$$

which gives

$$e_C(t) = E_{OS} \frac{K}{1 + K} (1 - e^{(1+K)t/RC}) \tag{7-2-2}$$

as a function of time. If $K \gg 1$, the voltage across C approaches E_{OS} exponentially with a time constant of RC/K. After a few reset cycles, the offset voltage is almost perfectly balanced out by the voltage across C, which is stored during the interval of amplification.

Since K is high, sudden offsets or input noise can easily drive the amplifier into nonlinearity in the effort to satisfy Eq. (7-2-2). To prevent saturation and the ensuing slow recovery, and to improve operation in other ways, the circuit in Fig. 7-13 was developed. The resistor R_2 controls the charging rate of C and limits the rate to a value that avoids amplifier saturation. Now C responds only to the low-frequency components of input offset; if it responds to input noise and the switch opens at times that are arbitrary with respect to the noise variations, then the voltage to which C charges has a random variation. Instead of the photoconductor, a fast reed switch S_3 is commonly used to give a very low switch resistance and lower switch offset voltage. During the amplifying phase of operation, S_3 is open and S_4 is closed, which establishes feedback through R_1 and R_2 to stabilize gain without interfering with the purpose of C.

The switches S_1 and S_2 in Fig. 7-13 are outwardly equivalent to S_1 in Fig. 7-1, but they were physically designed to have less thermal error.

One suitable application for the automatic-reset dc amplifier is in oscilloscopes, where reset can be accomplished during the time between individual traces of the pattern. Another application is in analog-to-digital converters, where the conversion is done periodically.

A commercial amplifier using the reset technique has the following per-

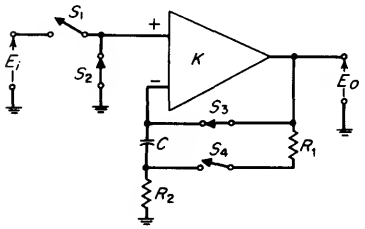


FIG 7-13 Modified reset amplifier in reset mode.

formance characteristics:

Noise (dc, 30 Hz) $< 1\ \mu$V peak-to-peak
Response time to 0.004 percent < 2 msec (gain $= 1$ and small signal levels)
Response time to 0.004 percent < 25 msec (gain $= 1$, $V_i = 10$ V) (limited by slew rate)
Offset drift $< 5\ \mu$V for 3 months
Offset stability with temperature $< 0.2\ \mu$V/°C (-20 to $+55$°C)
Leakage currents < 25 pA
Dynamic range (input voltage $= \pm 15$ V) (gain $= 1$)
Gain accuracy ≈ 0.001 percent of level
Input impedance $> 10^{10}\ \Omega$

A word of explanation is needed at this point in regard to response time. When an amplifier is used to measure the sudden application of a small dc signal, and one wishes to read the measurement with maximum accuracy as soon as possible, response time is a more convenient concept than frequency response. Such a situation occurs in low-level digital voltmeters with resolutions of five or six digits and reading rates of perhaps 15 sec^{-1}. In this case, it is quite probable that the output of the dc preamplifier must settle to within 0.004 percent of its final value within about 15 msec.

As implied in the performance characteristics above, the response time depends in large measure upon whether or not the preamplifier is saturated. If the amplifier is operating linearly, the response to a step function can be easily calculated from the transfer function. However, if a stage of the amplifier saturates because of a step-function input, the steady-state value of which would not cause steady-state saturation, the difficulty is that the saturating stage cannot supply the current required to charge a capacitive load at the desired rate of change of voltage. The rate at which the output of the amplifier can reach the correct level during saturation is called *dynamic slewing rate*. This matter is discussed more fully in Chap. 13, a general treatise on amplifier measurements.

7-3 Differential Amplifiers

Figure 7-14 shows an amplifier with two input terminals, neither one of which is connected to the system ground. If this is a differential amplifier, the useful (or differential) gain is

$$K = \frac{\Delta E_o}{\Delta(E_1 - E_2)} \tag{7-3-1}$$

and the common-mode gain

$$K_{CM} = \frac{\Delta E_o}{\Delta(E_1 + E_2)} \tag{7-3-2}$$

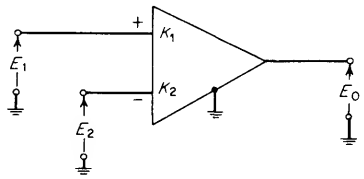


FIG 7-14 Elementary differential amplifier.

is made as small as possible. In other words, the noninverting gain $|K_1|$ is made nearly equal to the inverting gain $|K_2|$. Several differential amplifiers have been shown in diagrams earlier in the chapter, but the second input terminal has been used for convenience in a feedback scheme; the emphasis has not been upon amplifying a differential signal accurately in the presence of a common-mode signal.

However, the differential configuration *is* used for rejection of common-mode voltages. It is also useful for reduction and stabilization of input voltage offsets. For the latter purpose, the designer attempts to make the offsets of two nearly identical input devices affect the output in opposite directions and thereby cancel. The progress of integrated-circuit technology has played an important role in improving differential amplifiers because it has produced well-matched active devices in close proximity to one another so that the cancellation of environmental effects is more nearly perfect.

It is generally difficult to realize the desired overall characteristics with only one open-loop differential "block" of integrated circuitry, and so instrumentation amplifiers often utilize several interconnected blocks. One possible approach is shown in Fig. 7-15. In this circuit, A_1, A_2, and A_3 are integrated-circuit amplifier blocks with negligible input currents. All are inverting, as shown by the minus signs, and the corresponding voltage gains are $-K_1$, $-K_2$, and $-K_3$, which are all assumed to be of the

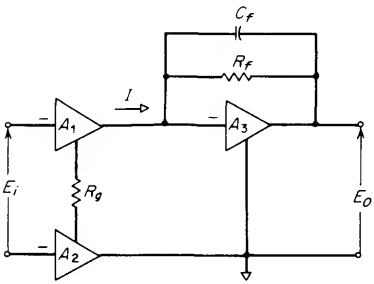


FIG 7-15 Differential amplifier with three integrated-circuit blocks.

order of 10^6. It is easy to derive the relationship

$$I = \frac{KE_i}{R_g(1 + K) + R_L} \tag{7-3-3}$$

where R_L is the equivalent load resistance on A_1. $K = K_1 = K_2$, and the other symbols are shown in Fig. 7-15. Since $K \gg 1$ and $R_L \ll KR_g$ because of the shunt feedback around A_3,

$$I \approx -\frac{E_i}{R_g} \tag{7-3-4}$$

The arrangement of A_1 and A_2 is called a *transconductance amplifier*, and the common-mode gain would be zero if the two blocks were perfectly balanced.

The transimpedance connection of A_3 responds to input current according to the equation $E_o = -IR_f$, and thus the gain of the whole amplifier is

$$K_d \approx \frac{E_o}{E_i} = \frac{R_f}{R_g} \tag{7-3-5}$$

The gain and bandwidth can be controlled independently since the gain may be varied by changing R_g and the bandwidth is determined primarily by R_fC_f.

A different closed-loop connection is shown in Fig. 7-16. This configuration uses a total of four amplifier blocks rather than three, but the requirements on each are not as severe, making the circuit more compatible with integrated-circuit technology. The operation is as follows: Blocks A_1 and A_2 respond only to the differential input signal E_i, which produces

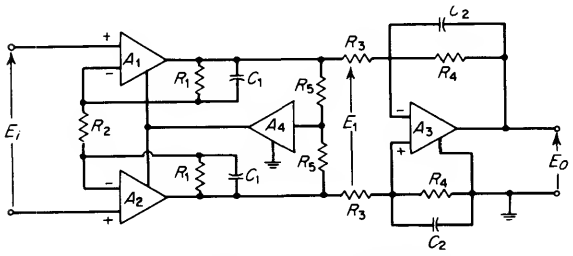


FIG 7-16 Improved differential amplifier with four integrated-circuit blocks.

a balanced output voltage

$$E_1 = \frac{(2R_1 + R_2)E_i}{R_2} \tag{7-3-6}$$

This signal is amplified a second time by A_3, according to the formula $E_o = -R_4E_1/R_3$, so that the overall voltage gain can be given as

$$\frac{E_o}{E_i} = \frac{(2R_1 + R_2)R_4}{R_2R_3} \tag{7-3-7}$$

While gain and bandwidth now depend upon a larger number of components, reasonable switching ease and stability are not hard to achieve. The function of A_4 is to reconstruct the common-mode voltage by means of summing resistors R_5 and to drive the power common and guard shields of A_1 and A_2 to prevent deterioration of the intrinsic CMR (common-mode rejection) by stray capacitances. In the configuration of Fig. 7-15, the CMR depends primarily upon the minimum values of the open-loop gains of A_1 and A_2, while in the configuration of Fig. 7-16, the CMR depends primarily upon the matching of the R_3's and R_4's to each other. While this may be construed as either an improvement or a disadvantage, depending upon circumstances, it is also possible to match the C_2's and achieve lower CMR at higher frequencies than is possible with the Fig. 7-15 circuit. Remember CMR $= K_{cm}/K_d$.

A comparison between the performance characteristics of differential amplifiers and chopper-stabilized amplifiers is made at the end of the chapter. Power-supply considerations limit the amplitude of common-mode voltage to about ± 10 V in differential amplifiers.

7-4 Chopper Amplifiers

Frequently the most troublesome problems of the dc amplifier are (1) the requirement of a finite input to bring the output to zero (zero offset), (2) a slow variation of the output when the input is constant (drift), and (3) an unpredictable current in the input leads (leakage current). One important technique for reducing these effects is to make the dc signal modulate an ac carrier so that amplification can be done without direct coupling, and then the amplified, modulated carrier is demodulated to regain a dc signal.

While this technique has been stated in general terms, the modulation method commonly used is to switch the input of an ac amplifier repetitively between the dc source and some reference potential such as ground. Figure 7-17 illustrates the method. The repetitive switches are called *choppers*. If the input chopper were ideal, the dc signal would be con-

verted into a square wave with amplitude equal to E_i and fundamental frequency f_c. The ac amplifier may amplify either the full square wave or else just f_c and partially the modulation sidebands. For qualitative understanding, it is easy to think of the output of the amplifier as a square

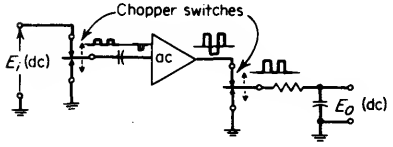


FIG 7-17 A basic chopper amplifier.

wave that is sampled synchronously by another chopper and then is filtered to produce a dc output.

Chopper Switches. The most commonly used chopper or modulating switch is basically a mechanical relay, driven magnetically. However, a good chopper is a very highly developed device that minimizes several severe problems. First, consider the choice of chopping frequency f_c. For two reasons it is desirable for f_c to be as high as possible: (1) The bandwidth of the dc system is seldom greater than $0.1f_c$, because of demodulation and filtering, and (2) many amplifying devices exhibit excess noise voltage that rises at a rate of approximately 3 dB/octave as frequency is decreased below about 500 Hz. The sampling theorem tells us that "in order to recover a continuous time series from its samples, except for a time delay, the sampling rate must be greater than twice the highest frequency in the continuous time series" [2]. In practice, 0.1 to $0.2f_c$ is a usable bandwidth. In addition, a time delay in transient response as great as $\frac{1}{2}f_c$ can exist [4].

Since the mechanical-chopper switches are driven magnetically, it is difficult to keep from inducing voltages in the switch conductors that appear as offset and noise after demodulation and can even cause currents to flow back into the source of the dc signal.

Jitter in the switching cycle and bouncing of the switch contacts are additional sources of noise. Also, the voltages across the thermal junctions within the choppers cannot be perfectly balanced to zero. Some thermal drift and offset remains. Still, mechanical choppers are used in amplifiers with bandwidths of about 10 Hz and noise referred to input of about 0.2 μV when used with signal sources having only a few ohms of internal resistance.

A conventional amplifier circuit with only one chopper switch (*SPDT*) is shown in Fig. 7-18. For simplicity, the chopper in this circuit is usually driven at line frequency, but it is then possible for line frequency interfer-

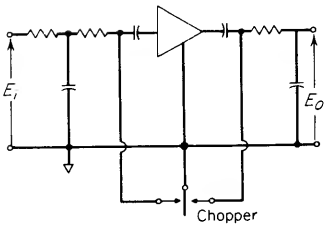


FIG 7-18 Direct-current amplifier with one chopper device for both modulation and demodulation.

ence and harmonics thereof, in either the input circuit or the ac amplifier, to produce an unpredictable dc output.

One improvement in the chopper amplifier for use with low-impedance sources and high common-mode voltages is to employ a full-wave system rather than the half-wave design described above. Such a configuration (Fig. 7-19) commonly makes use of a transformer to perform this function; instead of connecting the input signal to ground for one-half cycle, the input signal is inverted and used. Now the amplifier is connected to the signal nearly all the time and impedance adjusting is possible so that lower noise figures can be realized, but another precision component has been introduced which produces additional advantages and disadvantages.

One totally new advantage is the separation of the input and output ground paths, which results in improved performance where ground potential differences occur. The disadvantages include severe transformer design and shielding problems; frequency limitations; and if the isolation feature is used, the dependence upon the transformer losses as a limiting factor in accuracy since there cannot be any overall feedback to the isolated input except through another set of transformer windings and choppers.

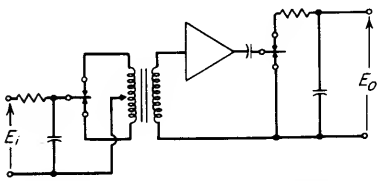


FIG 7-19 Direct-current amplifier with full-wave choppers and transformer isolation.

Other Choppers. Several nonmechanical devices are widely used as modulating switches to convert dc signals into ac signals in chopper amplifiers. Photoresistors (photoconductors) were first used successfully in

1957 as light-actuated switches in a stable dc microvoltmeter with 10 μV full scale [3]. The circuit of the microvoltmeter is basically similar to the one shown in Fig. 7-17, with much series feedback to stabilize the gain and increase the input resistance, which is greater than 100 MΩ. Photoconductors had long been used in crude switch applications, but early devices had far too high a generated photovoltaic potential and were far too slow in response time for use in chopper circuits at the microvolt level.

A carrier or chopping frequency of 50 Hz is used in the microvoltmeter described above. Direct-current amplifiers using higher chopping frequencies have been designed since 1958, but faster photoconductors having lower resistance with illumination (R_{ON}) would bring even further improvement in these amplifiers. Conductance is roughly proportional to light intensity on the photoconductive surface, but for practical reasons, R_{ON} is seldom less than 1,000Ω.

Transistors are excellent and fast switches for some chopper applications. The simplest circuit is shown in Fig. 7-20 with a single transistor used as a series switch. During the half cycle of the drive voltage E_d when the base of the p-n-p transistor is driven positive with respect to the emitter, very little current flows anywhere in the device, regardless of whether the input is positive or negative. On the other hand, when the base is driven negative, a current flows from emitter to base, and the emitter becomes a fraction of a volt positive with respect to base. Now, a collector current will flow if the input is made either positive or negative. As long as the collector current is restricted in magnitude, the collector-base junction will remain forward biased so that the net collector-to-emitter voltage is quite low. The main difficulty with this simple circuit is that an offset occurs in the E_o versus E_i curve, and the offset drifts with temperature.

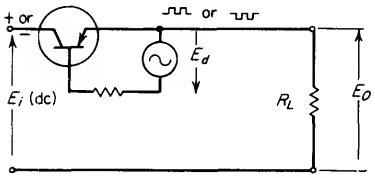


FIG 7-20 Transistor used as a chopper switch.

The offset error is reduced by inverting the transistor, which reverses the functions of emitter and collector. Even better reduction of offset is made by using *two* transistors in a circuit in which their offsets tend to cancel, as in Fig. 7-21.

In any chopper circuit, keeping the active chopper elements at constant temperature reduces drift. If two elements are used to cancel drift, it is important to keep the two at the same temperature, and a good way to prevent temperature differentials is to make both elements a part of the same integrated circuit. The most common chopper device of this sort is an integrated double-emitter bipolar transistor.

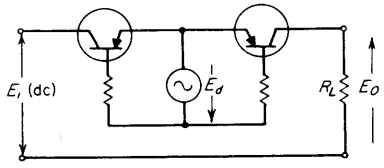


FIG 7-21 Bilateral transistor chopper.

The FET has virtue as a chopping switch because of its low offset voltage and low driving power. Some of the characteristics of chopping devices are compared in Table 7-1.

Another modulating device that is used in amplifiers requiring extremely high input impedance is the vibrating capacitor, the capacitance of which is made to vary periodically by mechanically oscillating the position of one of its plates [5]. The input voltage is applied to the capacitor through a high resistance so that the charge remains nearly constant over a cycle of vibration. The ac output voltage is proportional to charge multiplied by variation in capacitance, and currents as low as 10^{-16} A can be resolved. However, drift and noise are greater than in most low-impedance choppers. Although this type of chopper is well suited to resolving extremely low currents, it is not well suited to resolving extremely low voltages.

TABLE 7-1 Characteristics of Some Chopping Devices

Device	f_m, Hz	V_{OS}, V	I_{OS}, A	R_{ON}, Ω	R_{OFF}, Ω	V_{OFF}, V
Mechanical	10^2	10^{-6}	10^{-10}	10^{-1}	10^9	30
Photoresistive	3×10^2	10^{-6}	10^{-11}	10^4	10^6	50
Inverted Q†	10^6	10^{-3}	10^{-9}	25	10^8	5
Integrated double Q‡	10^6	10^{-4}	10^{-6}	50	10^8	20
FET	10^5	10^{-6}	10^{-9}	10^2	10^6	20

† Inverted transistor.
‡ Two transistors on one integrated-circuit chip.

CITED REFERENCES

1. Uzunoglu, Vasil: "Semiconductor Network Analysis and Design," McGraw-Hill Book Company, New York, 1964.
2. Nichols, M. H., and L. L. Rausch: "Radio-Telemetry," 2d ed., John Wiley & Sons, Inc., New York, 1956.
3. Cage, J. M.: An Increased-Sensitivity Micro Volt-ammeter Using a Photoconductive Chopper, *Hewlett-Packard J.*, vol. 9, no. 7, March, 1958.
4. Massey, W. S.: A Review of the Transistor Chopper, *Airpax Tech. J.*, vol. 1, no. 2, April, 1960.
5. Caldecourt, V. J.: Using a Vibrating Capacitor as an Electrometer Input, *Electronics*, vol. 35, Apr. 6, 1962.

CHAPTER EIGHT

VOLTAGE AND CURRENT MEASUREMENTS

From Notes by

Paul Baird, Arndt B. Bergh, Robert L. Dudley, William McCullough

Hewlett-Packard Company

and by

Charles O. Forge

Durrum Instruments
Palo Alto, California

Along with impedance and power, the basic variables measured electronically are voltage and current. Chapter 1 has discussed the system of units for these quantities and their standardization. It was shown that standardization requires extreme attention to accuracy and therefore special measurement and comparison techniques. While these techniques are basic and essential to the electronic arts, they will not be treated further in this book.

This chapter will treat several subjects relating to voltage and current measurement that do not logically fall elsewhere. First, the topic of dc

digital voltmeters (DVMs) deserves a place of its own, because these instruments have become extremely accurate and versatile, and fairly inexpensive, during recent years. They interface well with other digital instruments, including computers, and therefore are important in the growth of instrument systems (Chap. 18).

Then, the measurement of ac voltages of countless waveforms requires the thorough understanding of some technical principles. These will be developed.

The chapter will conclude with a brief discussion of some important instruments that measure either ac or dc currents in conductors, without breaking into these conductors or measuring the voltages across known impedances.

8-1 Introduction to DVMs

Digital voltmeters are measuring instruments that convert analog voltage signals into digital presentations. The digital presentation can take the form of a front-panel readout or an electrical digital output signal. As the name implies, any DVM is capable of measuring analog dc *voltages*. However, with the appropriate signal conditioner preceding the input of the DVM, many other quantities can be measured. Some of these are ac voltage, ohms, dc and ac current, temperature, and pressure, and there are many others. The common element in all these signal conditioners is a dc output voltage proportional to the level of the unknown quantity being measured. This dc output is then measured by the DVM, and appropriate annunciation in the digital presentation indicates the quantity being measured.

Digital voltmeters have made significant contributions in the field of electronic instrumentation since their introduction more than 15 years ago and are enjoying a growing popularity as bench instruments and systems components. Among the many factors that distinguish DVMs from other voltage measuring instruments are speed, automatic operation, and programability. Although there are instruments such as precision differential voltmeters that can be more accurate than most DVMs, few instruments bring to bear on the measurement problem the same combination of speed and accuracy as does the DVM.

Automatic operation and programability are features of many DVMs that make them useful in systems applications where the need is for great versatility in measurement capability, high speed, and computer controllability. Many of these features are expensive, however. In measurement situations where speed, convenience, and automatic operation are not high in priority, instruments of less cost than the precision DVM would be better solutions to the measurement problems.

There are several varieties of DVMs. They differ in the following ways:

1. Number of measurement ranges
2. Number of digits
3. Accuracy
4. Speed of reading
5. Normal-mode noise rejection
6. Common-mode noise rejection
7. Digital outputs of several types

The basic measurement range of most DVMs is either 1.00 or 10.00 V. However, with the appropriate preamplifier, measurements can be made to the nearest 0.1 μV. The number of digits commonly varies from three to six. Most DVMs have a certain amount of overrange capability, which is indicated by an additional *leading* digit, usually 1. This digit should not be confused with a full decade.

The accuracy of most DVMs is commensurate with their resolution; i.e., a three-digit instrument could hardly claim an accuracy better than ± 0.1 percent since this is the basic resolution of the instrument. Accuracy much poorer than ± 0.1 percent would be a poor compromise, too. The ultimate in accuracy for short periods of time in controlled environments can be as good as within an error of ± 0.0015 percent of reading or ± 0.0002 percent of full range in a six-digit instrument.

The maximum speed of reading and the period of digitizing are interrelated. Some DVMs are able to digitize in 1.0 msec (or even much shorter time intervals) and hence could make up to 1,000 readings per second. However, capacitance of the input, whether it is stray or part of a filter, may limit the useful measuring speed from inputs with high source impedances to 100 readings per second or below. Besides, it is impossible to follow the visual readout at high reading speed. At high reading speed, the difference between a DVM and an analog-to-digital converter may be as simple as the presence or absence of visual readout.

Noise rejection is a subject that will be treated in more detail later in this section. Normal-mode noise rejection is usually achieved through input filtering or through the use of the integration technique. Common-mode rejection is achieved through the use of guarding or other ways of obtaining differential response. Digital outputs are usually in the form of four-line BCD code, but there are 10-line outputs or single-line serial outputs.

Probably the greatest single distinction between types of DVMs is the method used in converting the analog dc signal into a digital presentation. During the short time that DVMs have existed, many different techniques have been developed by which this conversion takes place.

In general, these techniques can be divided into two main types: integrating and nonintegrating. The main nonintegrating types of DVM techniques are:

Potentiometric
  Servo
  Successive approximation
  Null balance
Ramp
  Linear
  Staircase

Those that are integrating in nature are:

Voltage-to-frequency converter
Potentiometric to integrating
Dual slope

8-2 *Nonintegrating Types of DVMs*

Potentiometric. Stripped of all complications, the operation of the potentiometric technique is similar to that of the differential voltmeter. Figure 8-1 shows this arrangement in its simplest form. The linear divider is adjusted until the null indicator shows equality of the input voltage and the output voltage of the divider. Assume that the internal reference is equal to +10 V and the divider is set to a ratio of 0.50020 at balance. The unknown voltage at the input is equal to

$$\begin{aligned} V_i &= V_{\text{ref}} \times \text{setting of divider} \\ &= +10.00 \times 0.50020 \\ &= +5.0020 \text{ V} \end{aligned} \tag{8-2-1}$$

The range of voltage that can be measured in this fashion is dependent upon the value of the reference voltage. In Eq. (8-2-1), it can be seen that this range could be increased by increasing the value of V_{ref} since the maximum setting of the divider is 1.00000. Further, the sensitivity

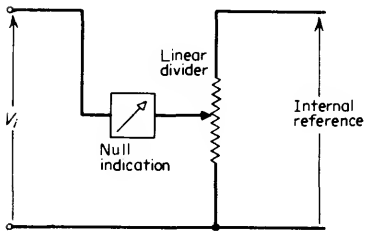


FIG 8-1 Simplified differential-voltmeter block diagram.

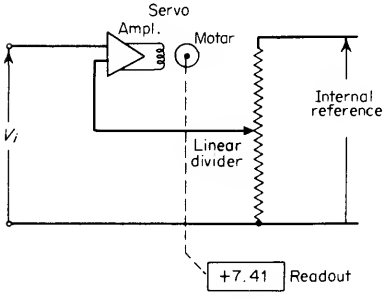


FIG 8-2 Simplified block diagram of servo DVM.

of the measurement is dependent upon the resolution of the linear divider. If it can divide with six-digit resolution, the resolution of the measurement is six digits and the maximum sensitivity is 10 μV. The sensitivity of this measurement is also dependent on the characteristic noise of the null indicator. These limitations can be partially overcome with the addition of an input amplifier and attenuator. The attenuator allows voltages higher than the internal reference to be measured, while the amplifier increases the sensitivity of the system and allows full-scale voltage ranges less than the internal reference. Voltages with negative polarity can be measured by reversing the connections between the internal reference and the linear divider.

Figure 8-2 shows the use of a *servo system* in automating this potentiometric technique. Here is a servo system composed of a differential amplifier, a motor, and the linear divider. The amplifier senses the polarity of the unbalance and drives the motor in such a direction as to reduce the unbalance to the level of some error signal. This level can be decreased by increasing the gain of the system—usually by increasing the gain of the amplifier. The accuracy of this system is dependent upon the internal reference, the linearity of the divider, friction, and the drift stability of the error amplifier. In this type of system, the divider is usually a multiturn potentiometer and is the main influence in determining the accuracy of the system. Attached to the shaft of the motor and divider is a mechanical readout which in effect indicates the position of the shaft of the divider. Because of the limitations in linearity of the potentiometer, the resolution of this system is usually a maximum of three digits.

This type of servo DVM is very economical in obtaining a digital readout of an analog input signal. Some of the disadvantages of this technique are its relatively slow speed and mechanical wear.

In the *successive-approximation technique*, the linear divider of the servo technique is replaced with a digital divider or digital-to-analog

converter and the servo motor is replaced with electronic logic. The readout in the successive-approximation technique is electronic and is usually driven by the logic of the system.

Figure 8-3 shows a simplified block diagram of a successive-approximation technique. This system is clocked (programmed) and usually follows a definite sequence of events. The digital-to-analog converter, under the control of the logic and sequencer, generates a set pattern of output voltages usually starting with the most significant digit of the measurement. As an example, assume that the input voltage is +3.7924 V. Further, assume that the code of the digital-to-analog converter for this particular example is 8 4 2 1, and the basic range of the instrument is 10.00 V. The sequence of events for this measurement is as follows:

1. Following the previous reading, the digital-to-analog converter is reset to 0 V. At the start of the next reading, the logic and sequencer cause the digital-to-analog converter to generate +8.000 V.

2. At this point in time, in synchronism with the logic, the input switch S_1 is connected to the output of the digital-to-analog converter. The capacitor C_1 charges to the 8.0000 V being generated by the converter.

3. The next event finds S_1 connected to the input voltage. The comparison amplifier now senses the direction of current flow at its input. If the input voltage is higher than that being generated by the digital-to-analog converter, the current will flow into the amplifier. However, if the input voltage is smaller than that being generated by the converter, the current will flow in the opposite direction.

4. In the example at hand, the comparison amplifier senses that the input voltage is smaller than the output of the digital-to-analog converter. A "high" decision is issued to the logic circuitry, and the 8.000-V condition in the converter is reset.

5. Next, the switch is returned to the output of the digital-to-analog converter, and the logic programs the converter to generate +4.000 V.

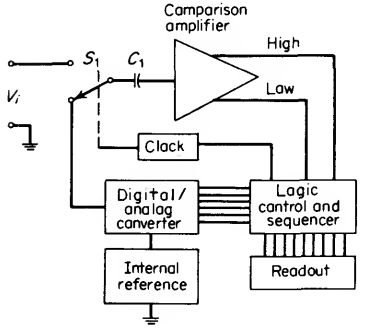


FIG 8-3 Simplified block diagram of the successive-approximation technique.

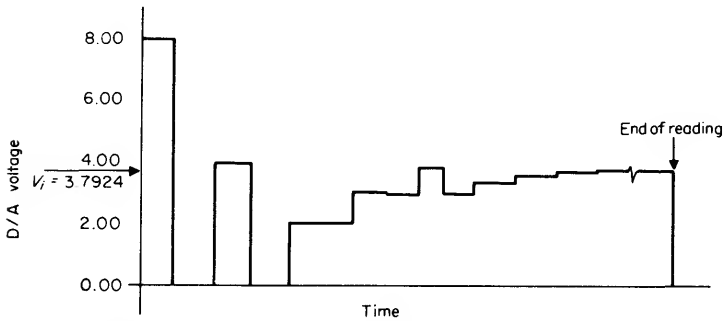


FIG 8-4 Digital-to-analog voltage sequence in successive-approximation measurement.

6. By comparing this voltage with the input, the same decision is made, since +4.000 V is still larger than the input voltage.

7. The digital-to-analog converter is next asked to generate +2.000 V. Capacitor C_1 charges to this level, and S_1 switches to the input voltage. This time, the current through C_1 flows opposite to the direction it had during the last two comparisons, and a "low" decision is issued to the logic.

8. Upon receipt of a low decision from the comparison amplifier, the voltage which was being generated by the digital-to-analog converter is retained, and the next incremental voltage is added to this amount. The special nature of the digital-to-analog converter makes this possible.

9. The next step has the digital-to-analog converter adding 1.00 V to what it has retained from previous decisions, in this case 2.000 V for a net output of +3.000 V.

10. Here again, a low decision is obtained, and the 3.000 V are retained in the digital-to-analog converter. The sequence now concentrates on the next digit in exactly the same fashion, and decisions as have been described above continue until the converter is generating +3.7924 V.

11. The last step in the measurement is the transfer of information from the sequencer into the readout.

Figure 8-4 shows the digital-to-analog voltage sequence in the measurement outlined above. As the figure illustrates, at each low decision after an incremental change in the digital-to-analog converter output voltage, this voltage approaches the value of the unknown voltage. The limit to how close in value these two voltages can become depends upon the level of noise in the input stages of the comparison amplifier and the stability of the input switch. These limiting factors usually determine the number of digits of resolution of the instrument.

The successive-approximation technique has been used in instruments with resolutions ranging from three to five significant digits. The speed of this technique depends upon the type of switches used in the digital-to-analog converter and comparison circuitry. If solid-state switches are used, speeds up to many thousands of readings per second can be achieved. This capability is needed in high-speed computerized measurement systems. If electromechanical switches are used, such as reeds or relays, speeds of a few readings per second can be achieved. The components in this technique that determine the basic accuracy are the internal reference supply and the digital-to-analog converter. These two items are common to many different types of DVMs and will be covered in detail in a later section.

The successive-approximation, or potentiometric, technique has been used nonautomatically for many years in the standards laboratory to achieve the highest possible accuracies while measuring dc voltage. The manual adjustment of a precision divider until a null is reached is quite simple, and yet it has stood the test of time in providing the ultimate in dc measurement accuracy. There are some practical considerations, however, which must be taken into account when applying this technique to the design of a DVM. Any noise that is in series with the input signal can cause incorrect decisions to be made by the comparison amplifier, which results in an incorrect measurement of the average value of the unknown. The usual solution to this problem is the addition of a low-pass filter at the input to pass dc information to the converter and attenuate ac components. Attendant on the use of an input filter is an increased response time of the instrument. Thus, in many practical situations, one may not be able to take advantage of the inherent high speed of the successive-approximation technique because of the use of the input filter to reduce the effects of noise.

The *null-balance technique* is virtually identical with that used in the successive-approximation technique except for the logic. In the null-balance technique, once a new voltage has been connected to its input, this instrument goes through the same steps in achieving a balance. However, once this balance is achieved, the digital-to-analog converter is not reset to zero as in the successive-approximation technique. The null-balance technique makes use of tracking logic. This means the digital-to-analog converter is able to follow the input voltage for changes below a certain selectable level. If the change in input voltage exceeds this level, the digital-to-analog converter is reset to zero and a measurement similar to that in the successive-approximation technique is made to achieve a new balance.

In many instances, the same instrument can be programmed to work in either the successive-approximation mode or the null-balance mode.

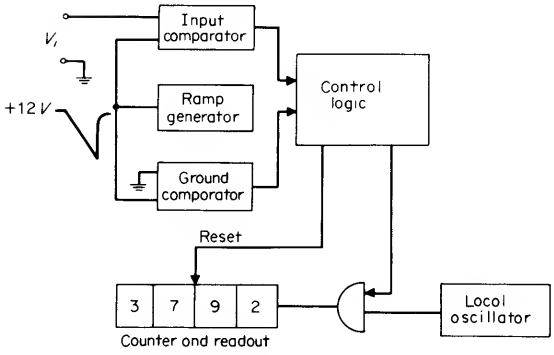


FIG 8-5 Simplified linear-ramp DVM.

One advantage of null balance is increased speed, since for small changes in the input voltage, a much smaller time is required to achieve a new balance.

Ramp Techniques. One consideration in the design of a DVM is the conversion of the analog dc signal into a quantity which is easily digitized. Two quantities that are relatively easy to digitize are time and frequency. The linear-ramp technique is essentially a voltage-to-time converter.

As the name of this technique implies, a linear ramp is used to convert an analog dc signal into a front-panel digital presentation. Figure 8-5 illustrates a simplified block diagram of a linear-ramp DVM. The heart of the system is the linear ramp itself, which in this case swings from +12 to −12 V. This voltage swing limits the basic measurement range of the instrument, which in this case is 10.00 V. The output of the linear ramp is connected to two comparators. These comparators are such that when one input becomes equal to the other, their output changes state. Thus the input comparator compares the ramp with the input signal, and when the two become equal, the output of this comparator changes state. Similarly, the ground comparator output changes state when the ramp passes through 0 V on its swing from +12 to −12 V.

The outputs of the two comparators are connected to logic circuitry, which controls the gate between a local oscillator and a counter. Prior to each measurement a reset pulse resets the counter to zero and sets the logic to a specific state. This logic is such that after reset, the first signal from either of the comparators will open the gate between the local oscillator and the counter, while the next signal will cause the gate to close. During the period the gate is open, the counter accumulates counts at a rate set by the local oscillator.

It is obvious that there is a direct relationship between the slope of the

ramp, the number of digits in the readout, and the frequency of the local oscillator. Assume that the slope of the ramp is 10 V/50 msec. That is, the ramp will travel from +10.00 V to 0/V in 50 msec. If a four-digit readout is desired, then the frequency of the local oscillator must be

$$f = \frac{10{,}000 \text{ pulses}}{0.05 \text{ sec}} = 200 \text{ kHz} \tag{8-2-2}$$

As Eq. (8-2-2) indicates, the resolution of the readout is directly proportional to the frequency of the local oscillator. As an example, assume that 3.792 V is applied to the input of the voltmeter in Fig. 8-5. When the ramp passes through a voltage equal to the input voltage, the logic causes the gate to open. When the ramp passes through 0 V, the logic causes the gate to close, and at the end of the ramp, a signal is generated which transfers the information in the counter to the front-panel readout. With a local oscillator frequency of 200 kHz and a ramp slope of 1.0 V/5 msec, the following calculations demonstrate numerically how the measurement evolves:

$$\begin{aligned} \text{Gate period} &= 3.792 \times 0.005 \\ &= 0.01896 \text{ sec} \\ \text{Number of pulses} &= 0.01896 \times 200{,}000 \\ &= 3{,}792 \text{ pulses} \end{aligned}$$

As shown above, all that remains is the proper location of the decimal point and indication of the quantity being measured. If the input voltage were reversed so that −3.792 V were to be measured, the first comparison would be supplied by the ground comparator. Remember, however, that the logic responds to the first signal from either comparator. The second signal in this case would come from the input comparator, and this would close the gate. Other logic has to observe which comparator changes state first in order to determine the polarity of the input signal and so indicate on the front panel.

It is obvious that the key elements in the accuracy of this technique are the linearity and absolute slope of the ramp and the frequency setting and stability of the oscillator. Offsets and drift in the two comparators are also important in determining the overall accuracy of the linear-ramp technique.

8-3 *Digital Voltmeters with Counting Circuitry*

The next technique to be discussed is classified as a ramp technique, but in many ways the staircase- or digital-ramp technique can be con-

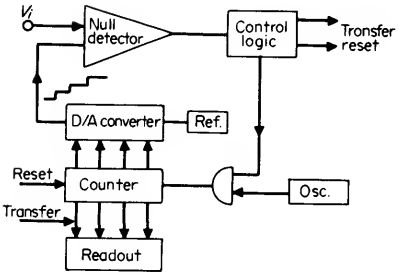


FIG 8-6 Simplified block diagram of staircase-ramp DVM.

sidered null balance or potentiometric in nature. This will become more obvious as we proceed with a description of its operation.

A simplified block diagram of this technique is shown in Fig. 8-6. The heart of the system is the digital-to-analog converter, which is the source of the staircase or digital ramp. The digital-to-analog converter is driven by the digital output of an electronic counter, which in turn is driven by an oscillator. At the beginning of a measuring sequence, the counter is reset to a zero-count condition by the control logic. Since the digital-to-analog converter is digitally slaved to the counter, it, too, assumes a zero state and hence generates zero output voltage.

Next in the measuring sequence is the opening of the gate, located between the oscillator and the counter, by the control logic of the system. The counter begins accumulating counts from the oscillator and causes the digital-to-analog converter to generate an output voltage equivalent to the instantaneous count. As an example, assume that the instrument under discussion is a three-digit voltmeter and is set to the 1.000-V range. This means that as each new count is entered in the counter, the output of the digital-to-analog converter increases by 1.0 mV. If the frequency of the oscillator is 1 kHz, the "slope" of the digital ramp is 1 V/sec. The ramp continues to build in value, 1 mV per step, until the null detector determines that the ramp has exceeded the input voltage with its last incremental increase. At this point the output of the null detector changes state. This is sensed by the control logic and the gate between the oscillator and the counter is closed. Following this gate closure, the transfer pulse is generated by the control logic and causes the count in the counter to be transferred to the front-panel readout.

It should be obvious now why the staircase-ramp DVM can also be considered null balance in nature in that a null is eventually generated at the input of the null detector. The reference supply and digital-to-analog converter are the primary factors in the accuracy of this technique. The operating speed is determined by the frequency of the oscillator and the number of digits the voltmeter has. Typically, this technique does not yield speeds faster than 10 readings per second.

Voltage-to-frequency Conversion. The next method to be discussed is the voltage-to-frequency conversion technique, and Fig. 8-7 shows a simplified block diagram. In this diagram, the input voltage causes a current to flow through R_1 into the summing junction of the operational amplifier. This current continues through C_1 and causes the output voltage of the operational amplifier to depart from 0 V. If the input voltage is positive, the direction of this change at the output is in a negative direction. If the input voltage is constant, the output moves at a constant rate in a negative direction.

When this voltage reaches a value equal to $-V$, the comparator shown in Fig. 8-7 changes state at its output. This triggers the pulse generator to inject a fixed amount of charge into the summing junction of the operational amplifier. The polarity of this charge is such that it tends to restore the output voltage of the operational amplifier to 0 V. The process described above continues to repeat itself, and a signal that looks very much like a sawtooth is generated at the output of the operational amplifier. If the input signal were doubled in value, the number of "teeth" in this output signal per unit time would also double. Coincident with each of these teeth is a pulse which passes through T_1 and on to the input of a control gate.

These pulses are allowed to enter the reversible counter when the gate is opened, and this opening of the gate is the beginning of the measurement cycle. The gate can remain open for any period of time, but typically this time is either 0.1 or 1.0 sec. During this period of time, the reversible counter totals these pulses. At the end of this period, the count stored in

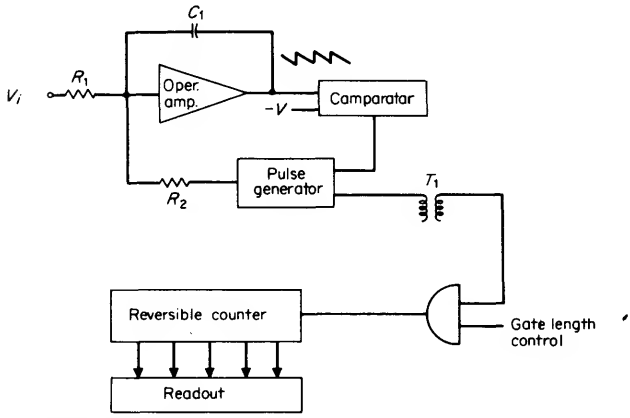


FIG 8-7 Simplified block diagram of voltage-to-frequency integrating DVM.

the reversible counter is transferred to the readout. The heart of this system is the circuitry that converts the input signal into the train of pulses to be counted. The accuracy of the system is totally dependent upon the magnitude and stability of the charge fed back to the summing junction of the operational amplifier by the pulse generator. The volt-time integral of this input signal and the total of the volt-time areas of the feedback pulses are kept in balance at the summing point by the rate at which the feedback pulses are generated.

Two of the more typical ways to design a precise pulse generator use either a transformer having a core of square-loop material or charge transfer from precision capacitors. In the case of the transformer, an excursion around the *BH* loop generates a precise amount of energy at a secondary winding, which is connected to the summing junction of the operational amplifier through a resistor. Where precision capacitors are used, they are allowed to charge to a known precise voltage. This charge is then transferred to the summing junction of the amplifier each time the output of the amplifier reaches $-V$.

The diagram in Fig. 8-7 shows only one comparator-pulse generator set. In actual practice there is another set to accommodate negative polarity signals at the input of the voltmeter. If the input voltage were negative, the output of the amplifier would go in a positive direction. Another comparator determines when this signal passes through $+V$ and triggers another pulse generator of opposite polarity. One important requirement is that both pulse generators produce identical amounts of charge each time they operate.

The need for a counter that has reversing capabilities should be discussed at this time. If an input signal were to change its polarity during a measuring period, those pulses that are accumulated following this must be subtracted from those that have been accumulated prior to this occurrence. If the counter were not able to reverse, there would be a folding over or *rectification* of the opposite polarity signals and a number larger than the actual average voltage would be indicated.

The most likely situation where the input voltage polarity will change is in the measurement of a low-level signal (a few tens of millivolts) in the presence of larger amounts of superimposed noise. The ability to reject such noise will be discussed in more detail later in this section. Because of the integrating capability of this system, the average value of the input signal is measured during the period of time the control gate is open.

Figure 8-8 shows the ideal transfer characteristics of a voltage-to-frequency converter. There should be a linear relationship between voltage and frequency until a certain saturation frequency is reached. The slope of this linear relationship must be the same for both positive

and negative voltages. The decision of whether a pulse should be added or subtracted is determined logically in the instrument; the source of the pulse—whether it is from the positive or negative comparator—is one input to the logic. The count present at any given time in the counter must also be considered. There are four separate conditions to consider:

Positive counting up
Positive counting down
Negative counting up
Negative counting down

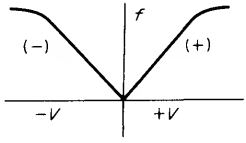


FIG 8-8 Transfer characteristics of voltage-to-frequency converter.

An example of these conditions is given in Fig. 8-9. In this example, it is assumed that an input voltage of 1.0 V causes the voltage-to-frequency converter to generate a pulse train with a frequency of 100,000 Hz. The input signal shown in Fig. 8-9*a* is rather complex with levels of +1.00, 0, and −1.00 V. Figure 8-9*b* shows the output frequency of the pulse generator responding to positive input signals, while Fig. 8-9*c* represents the output frequency of the generator responding to negative input voltages. Consistent with Fig. 8-8, the term *negative frequency* does not apply here.

The timing of the control gate is shown in Fig. 8-9*d*. Figure 8-9*e* shows the four conditions of count handling mentioned above. In the time period from $t = 0$ to $t = 0.2$ sec, the reversible counter is counting in a positive up direction. From $t = 0.3$ to $t = 0.5$ sec, the counter begins counting in a positive down direction. This continues until the net

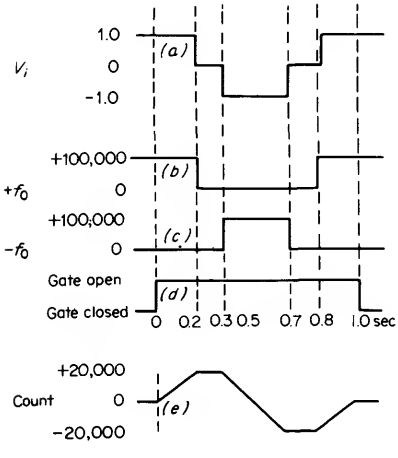


FIG 8-9 Integration of complex waveform.

count reaches zero, at $t = 0.5$ sec. At this point, the counter begins counting in a negative up direction until $t = 0.7$ sec. The counter now has a net of $-20{,}000$ counts entered. During the period from $t = 0.8$ sec to gate closure at $t = 1.0$ sec, the counter begins counting in a negative down direction. At the conclusion of the measurement, there is a net of zero counts in the counter. This is as it should be, since the integral of the input signal from $t = 0$ to $t = 1.0$ sec is 0. In this example, the time at which the count reached zero ($t = 0.5$ sec) was indicated to the control logic in order to reverse the state of the counter.

As was mentioned above, the accuracy of the voltage-to-frequency converter–integrating DVM is dependent on the precision of the charge that is fed back with each pulse and the linearity of the relationship between voltage and frequency. Typically, the highest order of accuracy obtainable in using this technique is to within 0.01 percent. The reading speed is dependent upon the number of digits in the measurement and the maximum rate of the voltage-to-frequency converter. If this rate is 100,000 pulses for an input voltage of 1.00 V, a five-digit reading would require 1.00 sec, while a three-digit reading would require only 10 msec.

The operating speed of an integrating DVM with voltage-to-frequency conversion can be increased by increasing the upper operating frequency of the converter. This usually results in decreased accuracy and added cost. A new technique has recently been introduced which uses voltage-to-frequency conversion and provides five-digit resolution at a reading rate approaching 50 sec^{-1}. Figure 8-10 shows a simplified block diagram of the interpolating-integrating DVM for this technique. For the first 16.67 msec of its operating cycle, it is very similar to the voltage-to-frequency integrating DVM described above. However, during this period the pulses generated are directed to the 100s decade: Each pulse is equivalent to 100 counts. At the end of the 16.67 msec period, there may be some charge remaining on the integrating capacitor C_1. At this point, S_1 switches to an internal reference whose polarity tends to reduce the voltage on C_1. At the same time, the gate feeding the 1s decade is opened and passes pulses at a 60-kHz rate. This continues until the zero comparator detects 0 V at the output of the operational amplifier. At this comparison, the gate passing 60 kHz closes and the reading is completed.

The amplitude of the internal reference is such that it can remove all the charge from a fully charged C_1 in the time it takes to enter 100 counts into the last two decades at a 60-kHz rate. Since the internal reference is a fixed value, the rate at which this charge is removed is also constant. This type of voltmeter has the same basic accuracy limitations as does the noninterpolating voltage-to-frequency DVM. The pulse generator must

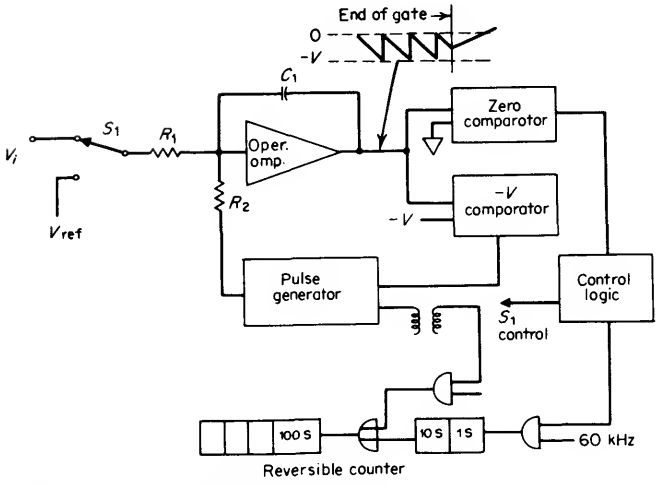


FIG 8-10 Interpolating-integrating DVM.

generate precise amounts of charge per operation, and this amount must be constant as the rate of operation changes. The determination of the last two digits of the reading need have an accuracy of only 1 percent to be consistent with their importance.

Potentiometric-Integrating DVM. Techniques discussed to this point are either potentiometric or integrating in nature. It was pointed out that the potentiometric DVM offers the highest accuracy, while the integrating type provides inherent rejection of noise in series with the signal to be measured. The next technique to be discussed is one that combines the potentiometric and integrating schemes and takes advantage of their best features.

A simplified block diagram of this DVM is shown in Fig. 8-11. Each measurement is composed of two separate and distinct sample periods. During the first sample period, a straightforward voltage-to-frequency measurement of the unknown is made. At this time in the measurement cycle, the digital-to-analog converter is generating 0 V, and only the unknown signal is impressed at the input of the voltage-to-current converter. The resulting current flows into the current-to-frequency converter, which generates a train of pulses, the repetition frequency of which is proportional to the instantaneous value of the input voltage.

These pulses are fed into the 100s decade of the reversible counter during this period. At the end of the period, the count in the counter

should equal the unknown voltage. At this point in the measurement cycle, the information in the counter is transferred to the digital-to-analog converter. The digital-to-analog converter is now called on to generate a voltage exactly equal to that represented by the count in the reversible counter. In the process of this transfer, the information is retained in the counter. The input to the voltage-to-current converter is now the difference between the input voltage and that voltage which the digital-to-analog converter is generating. This difference should be fairly small, but typically is not zero because of errors in the initial voltage-to-frequency conversion and the reduced resolution of this initial measurement.

At this point, the second sample period begins. The voltage-to-frequency converter (combined voltage-to-current and current-to-frequency converters) now generates a train of pulses, the frequency of which is proportional to the instantaneous value of the difference between the input signal and the output of the digital-to-analog converter. These pulses are now entered into the 1s decade of the reversible counter. Any carry pulses (every hundredth pulse) is fed to the 100s decade. At the end of this second sample period, the information in the reversible counter is transferred to the front-panel readout of the voltmeter.

The accuracy of this system is primarily dependent upon the reference supply and the digital-to-analog converter. The accuracy is good, and besides, the integrating capability of the voltage-to-frequency converter provides rejection of noise superimposed upon the input signal. One distinct advantage of this system is the reduced importance of the accuracy of the voltage-to-frequency converter. The following example will help to illustrate this.

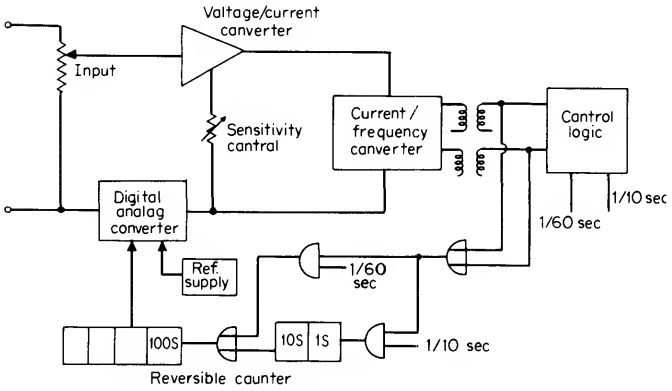


FIG 8-11 Potentiometric-integrating DVM.

Example

Assume there is no error in the digital-to-analog converter, while the voltage-to-frequency converter is in error ±0.5 percent of level. Assume the input voltage is exactly 1.000000 V. Because of the voltage-to-frequency error, the number of counts that is entered during the first sample period is 1,000 × 1.005 = 1,005 counts. This causes the digital-to-analog converter to generate 1.005 V during the second sample period.

Because of this difference, 500 counts should be subtracted during the second sample period. However, because of the voltage-to-frequency error, 503 counts will actually be subtracted. The net result is

$$\begin{array}{r} 1.005 \\ -503 \\ \hline 0.99997\ \text{V} \end{array}$$

As this example shows, even an error of 0.5 percent in the voltage-to-frequency converter produces only a 0.003 percent error in the final reading. Actually, the error of the voltage-to-frequency converter should be squared to determine its effect on the overall accuracy. Hence, it contributes an error of only 0.005 × 0.005 = 0.000025, or 0.0025 percent. In actual practice, the accuracy of the voltage-to-frequency converter is typically better than ±0.1% of reading so that its effect on a five-digit reading is negligible. This reduced importance of voltage-to-frequency accuracy is the reason the first two decades of the reversible counter are bypassed during the first sample period. Not only is the resolution of the first three digits sufficient for this sample, but bypassing the first two decades also speeds up the reading. By using a voltage-to-frequency converter with a maximum rate of 100 kHz, the first three decades can be easily filled in 16.67 msec. However, if all five decades were used, a 6-MHz voltage-to-frequency converter would have to be used, or a period of 1 sec would be required.

Because counts are fed into different decades during the measurement cycle, the sensitivity of the voltage-to-current converter has to be changed. This is accomplished by the variable resistor shown below the operational amplifier in Fig. 8-11. The use of this means of controlling the sensitivity allows the basic sensitivity of the current-to-frequency converter to remain unchanged during a measurement cycle.

Dual-slope Integration. A recent innovation has greatly simplified the conversion of an analog signal into a digital presentation. This is called the *dual-slope integration* technique, and a simplified block diagram is shown in Fig. 8-12. Basically, a DVM using dual-slope integration is a voltage-to-time converter, but not in the same sense as the linear ramp discussed earlier. The interpolation used in the interpolating-integrating DVM is dual slope in nature.

The input signal in Fig. 8-12 is integrated for a fixed interval of time. During this time, charge builds up on C_1. The next step is to measure the

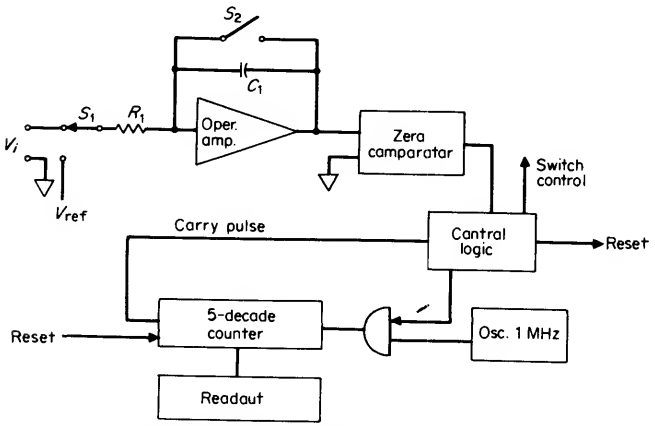


FIG 8-12 Dual-slope integrating DVM.

time it takes to discharge this capacitor, with the use of a fixed reference current. During the integration period, the rate of charge buildup on the capacitor is proportional to the level of the input signal. However, during the discharge cycle, the rate of discharge is fixed.

In the simplified block diagram shown in Fig. 8-12, the period of integration is determined by a 1-MHz oscillator and the counter. Just prior to a measurement, the counter is reset to zero. At the beginning of the measurement cycle, the gate between the oscillator and the counter is opened by the control logic, and at the same time S_2 opens. Even though S_1 has been connected to the unknown signal, the charge on the capacitor at the beginning of the measurement period is zero with S_2 closed. As soon as S_2 opens, charge begins accumulating on C_1 at a rate proportional to the input signal. This process continues until a carry pulse is generated by the counter. The counter resets to zero as the carry is generated. Since this is a five-decade counter and the counting rate is 1 MHz, the carry pulse is generated 100 msec after the opening of the gate. The carry pulse input to the control logic causes S_1 to switch to the V_{ref} position. The counter continues to accumulate pulses from the oscillator. The polarity of the reference voltage is opposite to that of the input signal, so that charge accumulated on C_1 during the integrating period is now removed by the reference signal. As soon as all the charge is removed, the output of the operational amplifier is at 0 V and this condition is detected by the zero comparator. This comparison signals the control logic to close the gate between the 1-MHz oscillator and the five-decade counter.

Some of the waveforms associated with dual-slope integration are

found in Fig. 8-13. The input waveform is shown in Fig. 8-13*a*. The solid line represents an input voltage of 1.000 V, while the dashed line represents an input voltage of 0.5000 V. Figure 8-13*b* indicates the waveform at the output of the operational amplifier. The slope of this voltage during the 100-msec integrating period is shown to be proportional to the level of the input signal. During the period following this integration interval, the slope of the output voltage is seen to be constant in both cases as it is returned to 0 V. When the input voltage is 0.5000 V, it takes only half the time to reach zero that it does with 1.000 V—hence there is voltage-to-time conversion.

Figure 8-13*c* shows the instantaneous count present in the reversible counter. During the integration period, the count steadily approaches 100,000 at a rate of 1 million counts per second. Since it is only a five-decade counter, it automatically resets itself to zero on reaching 100,000 counts and continues counting at the same rate until zero comparison is made at the output of the operational amplifier. As Fig. 8-13*c* shows, there are 50,000 counts present when zero is reached for $V_i = 0.5000$, while there are 100,000 counts entered for $V_i = 1.0000$ V.

The dual-slope technique puts the primary responsibility for accuracy on the internal reference voltage. The value of R_1 or of C_1 is not significant in determining accuracy. It is only required that they remain stable during the measuring period. The frequency of the oscillator, f, is similarly not important, provided it is constant. However, if line-related frequencies are to be integrated to zero by this technique, the period generated by the oscillator and counter should be some multiple of the period of the line frequency.

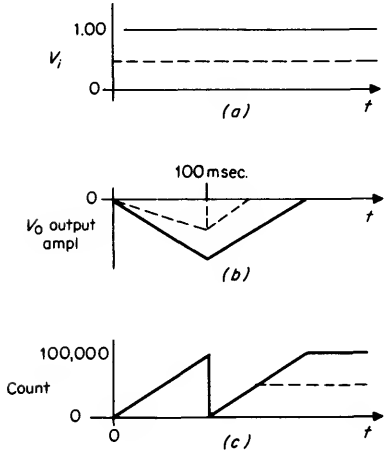


FIG 8-13 Dual-slope waveforms.

The reason that the values of R_1, C_1, and the frequency are not important in determining the accuracy is that they are used during both the integration period and the measuring period. The following expressions help demonstrate this:

$$V_{0\text{ op amp}} = \frac{1}{C_1}\frac{V_i}{R_1}T_{\text{integrate}} = \frac{1}{C_1}\frac{V_{\text{ref}}}{R_1}T_{\text{discharge}} \tag{8-3-1}$$

$$T_{\text{integrate}} = \frac{10^5}{f} \quad \text{and} \quad T_{\text{discharge}} = \frac{\text{accumulated counts}}{f}$$

$$\text{Accumulated counts} = \frac{V_i}{V_{\text{ref}}} \times 10^5 \tag{8-3-2}$$

Equation (8-3-2) shows that accuracy depends only on the internal reference voltage.

One comment can be safely made regarding the various techniques used to convert an analog signal into a digital output or presentation—Even though it would appear that the concerted efforts of industry over the past 15 years have discovered many techniques to make this conversion, the next clever technique may be described in the technical journal you read tomorrow.

8-4 *Normal-mode Rejection*

During the discussion of the theory of operation of integrating DVMs, reference was made to their ability to reject signals that are in series with the input signal. These signals are called *normal-mode* or *superimposed noise signals*. This type of signal can be caused by electromagnetic pickup in the input leads to the DVM, or by an inherent component of the unknown signal, i.e., ripple at the output of a power supply. Typically, the frequency of these signals is an integer multiple of the power-line frequency, but in practice any type of disturbance may be present.

There are many applications where fast readings are required to characterize the unknown signal adequately. The removal of superimposed high-frequency components would be unacceptable. However, in the following discussion, assume that only the average value of a signal is desired.

In the description of the successive-approximation technique above, it was observed that the reading is achieved only after a number of successful decisions by the logic circuitry. Any incorrect decision along the way can cause a completely erroneous reading of the unknown signal. Superimposed noise is the most common cause for bad decisions of this kind.

If the dc value of the unknown is somewhat above 8.000 V in value but noise causes it to drop just below this level when the digital-to-analog feedback is generating 8.000 V, the reading will be incorrect.

It is obvious that the severity of such errors is more or less proportional to the amount of noise present. The same is true of the linear- and digital-ramp techniques. Superimposed noise can cause bad comparisons between the input signal and the internally generated ramp. The usual solution to such problems is the inclusion of a low-pass filter at the input of the DVM. The presence of such a filter does reduce the amount of noise presented to the digitizing circuitry, but it also increases the time of response of such instruments. Any filter that reduces the amount of noise significantly has a typical settling time on the order of 500 msec to 1.0 sec. This settling time has to be added to the digitizing time of the instrument in determining the true reading rate.

All the integrating DVMs discussed above employ techniques that make use of a specific period of time called either a *gate length* or an *integrating period.* During this period of time, the signal present at the input of the DVM is truly integrated and the average signal present is indicated on the front panel. A good example of this is the waveform used in Fig. 8-9. Although this example was used in conjunction with the technique using voltage-to-frequency conversion, the same would be true of the dual-slope technique.

Assume the input to an integrating DVM can be presented by

$$v(t) = V_1 \sin \omega t \tag{8-4-1}$$

In this example, it is assumed that there is no dc component present in the input signal. In developing this general case, it will be assumed that the integration interval begins at $t - t_1$ and ends at $t - t_1 + T$, where T is the period of integration or the gate length. The average voltage during this integration period is

$$\begin{aligned} V_{av} &= \frac{V_1}{T} \int_{t_1}^{t_1+T} \sin \omega t \, dt \\ &= -\frac{V_1}{\omega T} \cos \omega t \Big]_{t=t_1}^{t=t_1+T} \\ &= -\frac{V_1}{\omega T} [\cos \omega(t_1 + T) - \cos \omega t_1] \end{aligned} \tag{8-4-2}$$

By expansion of Eq. (8-4-2), the average value is

$$V_{av} = -\frac{V_1}{\omega T} [-2 \sin \tfrac{1}{2} (2\omega t_1 + \omega T) \sin \tfrac{1}{2} (\omega T)] \tag{8-4-3}$$

where V_{av} can be maximized by choosing t_1 so that

$$\sin \tfrac{1}{2}(2\omega t_1 + \omega T) = 1 \tag{8-4-4}$$

Substituting Eq. (8-4-4) in Eq. (8-4-3) gives the expression

$$V_{av}(\max) = \frac{2V_1}{\omega T} \sin \frac{1}{2} \omega T \Big]_{\omega = 2\pi f}$$

$$= \frac{V_1}{\pi f T} \sin \pi f T \tag{8-4-5}$$

As the frequency of the superimposed noise approaches zero, V_{av} (max) approaches V_1. In order to develop an expression for attenuation as a function of frequency, it is necessary to establish the ratio of the value of the signal at 0 Hz to its value at a specific frequency.

$$\left| \frac{V_1}{(V_1/\pi f T) \sin \pi f T} \right| = \frac{\pi f T}{\sin \pi f T} \tag{8-4-6}$$

A plot of Eq. (8-4-6) is shown in Fig. 8-14. As seen in this figure, infinite cusps of rejection occur at intervals determined by $fT = n$, where $n = 1, 2, \ldots, K$. In this instance, the gate length is 1.0 sec.

By the appropriate choice of the integration interval, maximum rejection of line-related frequencies can be obtained. This is why many gate lengths are 16.67 msec, 100 msec, or other multiples of 16.67 msec in length. For use in Europe where the predominant line frequency is 50 Hz, 20.0 msec gate lengths should be used in place of 16.67 msec.

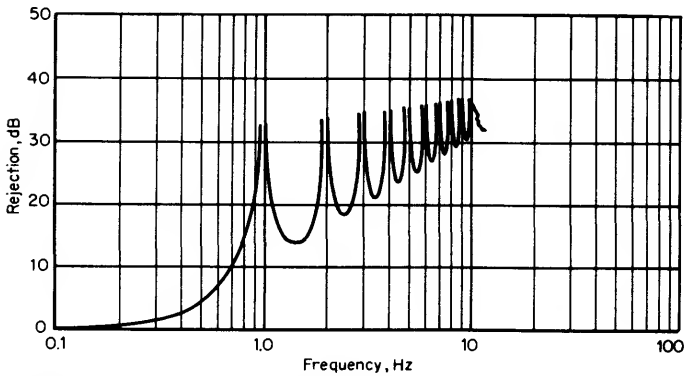


FIG 8-14 Normal-mode rejection characteristics of integrating DVM.

8-5 Common-Mode Rejection[1]

Another type of disturbance which can cause errors in measurements is made by common-mode signals. In Fig. 8-15, V_3 and V_4 are common-mode signals. In this figure, V_2 is a normal-mode signal, while V_1 is the signal to be measured. There are many sources of common-mode signals. If a floating measurement is to be made, the voltage on which the input of the voltmeter is floating is a common-mode signal. These signals can also be produced by ground currents, particularly where the source of the signal to be measured is some distance from the instrument making the measurement. In such a situation there are usually grounds used at each point, and the separation of these points usually means they are not at the same potential. The resistances R_1 and R_2 in Fig. 8-15 represent high- and low-side resistance in the measuring circuit.

Figure 8-16 demonstrates the practical situation of a load-cell measurement and the cell's equivalent circuit. A load cell is composed of four strain gages connected in a bridge arrangement with an isolated input and output. Figure 8-16*b* shows the equivalent circuit of the load cell in Fig. 8-16*a*.

All the other components shown in Fig. 8-15 are part of the measuring instrument. The resistance R_3 represents the input resistance of the voltmeter, while R_4 and C_4 represent leakages between the low terminal of the instrument and power ground, and R_5 and C_5 represent leakages between high and power ground. In most measuring systems, the impedance determined by the parallel combination of R_5 and C_5 is much larger in value at all frequencies than is the impedance of the parallel combination of R_4 and C_4. The reason for this is the fact that the high side is physically

[1] This discussion applies to instruments with floating, single-ended inputs.

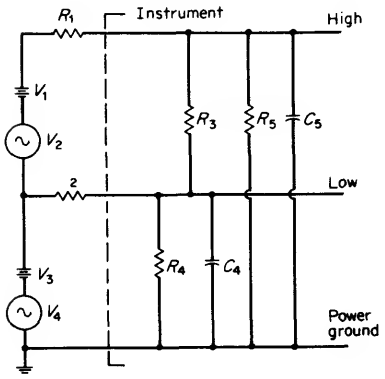


FIG 8-15 Noise signals and configuration of floating input.

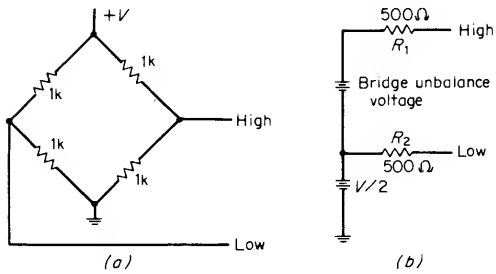


FIG 8-16 Load cell and equivalent circuit.

a wire or narrow conductor while the low side is a plane or large metal area in close proximity to the power ground of the measuring system, which is also a large surface. Because of this, the path for current through R_1 and the parallel combination of R_5 and C_5 will be neglected.

The disturbing fact about these paths is that current flowing through them will tend to generate a voltage that is in series with the signal to be measured. The most troublesome path of this nature is through R_2 and the parallel combination of R_4 and C_4. This current, of course, is due to the common-mode voltages V_3 and V_4. Common-mode rejection is related to the ability to reduce the voltage developed across R_2. Given a certain value of R_2 (taken nominally to be 1 kΩ in industry), the only way to increase this rejection is to increase the value of Z_4 (the parallel combination of R_4 and C_4).

In a well-designed floating instrument, R_4 may be as high as 10^9 Ω and C_4 may be as large as 2,500 pF. These values lead to the following CMRs:

$$\text{Direct current: } CMR = -20 \log \frac{10^9}{10^3} = -120 \text{ dB}$$

$$\text{Alternating current: } CMR = -20 \log \frac{1/2\pi fc}{10^3}\bigg|_{f=60 \text{ Hz}}$$

$$= -20 \log \frac{10^6}{10^3} = -60 \text{ dB}$$

In the above example, a dc common-mode signal of 100 V would develop 100 μV across R_2, while an ac common-mode signal of 20 V at 60 Hz would develop 20 mV across R_2. One can also think of CMR as a reduction in the amount of a common-mode signal converted into a normal-mode signal across R_2.

In many instances, the 100-μV dc signal or 20-mV ac signal would be intolerable in a measurement. This is particularly true if low-level

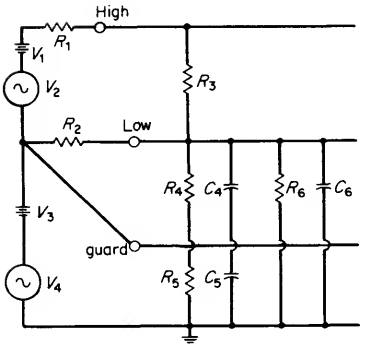


FIG 8-17 Configuration of floating and guarded input.

dc measurements are being made, as would be the case in many load-cell applications. A technique called guarding has been incorporated into many DVMs to increase their ability to reduce and eliminate the effects of common-mode signals. In its simplest form, a guard is a sheet-metal box surrounding the circuitry associated with low and is insulated from both low and power ground. A terminal at the front panel makes this "box" available to the circuit under measurement.

Figure 8-17 shows the application of a guard to the measurement situation shown in Fig. 8-15. In Fig. 8-17, Z_4 and Z_5 have the same magnitude as Z_4 in Fig. 8-15; i.e., the resistance is 10^9 Ω and the capacitance is 2,500 pF. However, R_6 in Fig. 8-17 is typically greater than 10^{11} Ω, while C_6 is less than 2.5 pF.

The resistance R_6 and capacitance C_6 do not physically exist; they cannot be measured directly, but represent what leakage remains "through" the guard when it is connected as shown in Fig. 8-17. In other words, the proper use of this guard considerably lowers the effective leakage between Lo and power ground; that remaining is indicated by the components of Z_6. The guard is actually driven by the common-mode signal, as is Lo; hence there is virtually no current through Z_4. The bottom side of Z_4 is essentially "bootstrapped" to low because the guard is driven by the same common-mode signal as low.

The CMR for the guarded circuit is:

$$\text{Direct current: } CMR = -20 \log \frac{10^{11}}{10^3} = -160 \text{ dB}$$

$$\text{Alternating current: } CMR = -20 \log \frac{1/(2\pi 60 \times 2.5 \times 10^{-12})}{10^3}$$

$$= -20 \log \frac{10^9}{10^3} = -120 \text{ dB}$$

This shows a considerable improvement over the unguarded situation. Now the 100-V dc common-mode signal generates only 1 μV of normal-mode signal, while the 20-V ac common-mode signal generates only 20 μV of ac normal-mode signal.

Effective CMR. Effective CMR is a concept that combines the effects of normal-mode rejection and CMR. It relates the effects of a common-mode signal to the readout of the instrument rather than to the input. Thus, if a certain instrument has a CMR of -120 dB at 60 Hz and has 50 dB of normal-mode rejection at the same frequency owing to an input filter, its effective CMR would be -170 dB at 60 Hz. Figure 8-18*a* shows a plot of the CMR of a guarded DVM with 160 dB of rejection at dc and 120 dB at 60 Hz. If this instrument is an integrating DVM with a guard, its effective CMR would be that shown in Fig. 8-18*b*. In this plot, it was assumed the gate length was 0.100 sec. As Fig. 8-18*b* shows, the effective CMR ratio is never lower than -145 dB, even as the frequency increases above 60 Hz.

8-6 *Principles of AC Voltage Measurements*

Most ac measurements are made with ac-to-dc converters, which produce a dc current proportional to the ac input being measured and use

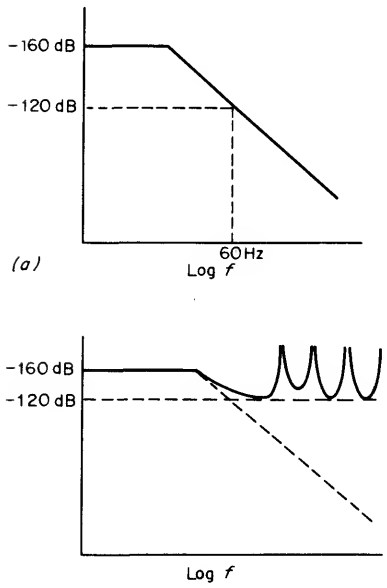


FIG 8-18 Effective CMR of integrating guarded DVM.

this current for either meter deflection or application to the dc circuitry of a digital or analog multimeter. Converting the signal to dc as soon as feasible minimizes the serious errors which otherwise could result from frequency-selective circuits.

Most ac voltmeters are classified into three broad types: rms-responding, peak-responding, and average-responding. Those that are average responding and peak responding are generally calibrated to read the rms value of a sine wave. Only a small minority of voltmeters available today are true rms-responding instruments, but it is expected that in the future many more true-rms voltmeters will be introduced.

Voltmeters are ordinarily calibrated in rms volts because that is the equivalent to a dc voltage which generates the same amount of heat power in a resistive load that the ac voltage does. For this reason, rms voltage is synonymous with effective voltage, and the term is used predominately in discussions of ac voltage without referring to the term *rms*.

The rms value of a waveform is defined as the square root of the average of the squares of the quantities being measured. Theoretically, rms value can be found by measuring the voltage at equal intervals along the waveform for one cycle, squaring the numerical value of the voltage at each point, finding the average value of all the squared terms, and then taking the square root of the average value. If there were n discrete values V_k in a series of measurements, this could be expressed as

$$V_{\text{rms}} = \left(\frac{1}{n}\sum_{k=1}^{n} V_k^2\right)^{1/2} \qquad \text{as } n \to \infty \tag{8-6-1}$$

If the quantity being measured is a continuous function of time, such as would normally be found in a voltage waveform, the summation process is replaced by integration. Then the rms value is expressed as

$$V_{\text{rms}} = \left(\frac{1}{T}\int_0^T v^2\,dt\right)^{1/2} \tag{8-6-2}$$

where the measurement is carried out through the interval from 0 to T. From this, the rms value of one-half cycle of a sine wave is found.

$$V_{\text{rms}} = \left(\frac{1}{\pi}\int_0^{\pi} (V_{\max}\sin\theta)^2\,d\theta\right)^{1/2} = (\tfrac{1}{2}(V_{\max})^2)^{1/2}$$

$$= 0.707V_{\max} \tag{8-6-3}$$

The *average* value of an ac voltage is simply the average of the voltage values measured point by point along the waveform; if there were n

discrete values within this period, this would be described as

$$V_{av} = \frac{1}{n} \sum_{k=1}^{n} V_k \tag{8-6-4}$$

If the quantity varies continuously, the average value is defined mathematically as

$$V_{av} = \frac{1}{T} \int_0^T v \, dt \tag{8-6-5}$$

It should be noted that in an analog-metered instrument, the averaging is performed by the inertia of the meter movement, while in a digital instrument, this process is generally performed by a low-pass filter.

The average value of one-half cycle of a sine wave is

$$V_{av} = \frac{1}{\pi} \int_0^{\pi} V_{max} \sin \theta \, d\theta = \frac{2}{\pi} V_{max} = 0.636 V_{max} \tag{8-6-6}$$

The average value of an integer number of cycles of a sine wave really is zero because the waveform has equal positive and negative values when it is averaged for a complete cycle. However, in the case of *average-responding* voltmeters, the average value of the sine wave is taken to mean the average rectified value or the average without regard to polarity.

The use of average-responding rather than rms-responding voltmeters is primarily a result of simpler construction and lower cost. It is justified by the wide use of sine waves in electronic measurements. In calibrating an average-responding meter, a pure sine wave with an rms amplitude of 1 V can be applied to the meter and the resulting pointer deflection on the scale is adjusted to read 1 V. This is done by applying a constant of proportionality to the meter called the *form factor*, which is defined as follows:

$$\text{Form factor} = \frac{\text{rms or effective value}}{\text{average value}}$$

The form factor of a sine wave is calculated as follows:

$$\text{Form factor} = \frac{V_{rms}}{V_{av}} = \frac{0.707 V_{max}}{0.637 V_{max}} = 1.11 \tag{8-6-7}$$

To determine the average value of a wave with the use of an average-responding voltmeter, simply divide the reading on the voltmeter by 1.11. To determine the peak value of a sine wave, multiply the reading on the voltmeter by 1.414. Obviously, to get the peak-to-peak value of a sine wave, one can multiply the reading on an average-responding voltmeter by 2.828.

8-7 *Average-responding Detectors*

A simplified version of the circuit used in a typical average-responding voltmeter is shown in Fig. 8-19. The applied waveform is amplified in a high-gain stabilized amplifier to a reasonably high level and then rectified and fed to a dc milliammeter calibrated in terms of the rms input voltage. In a digital multimeter, the rectified current is filtered and then applied to the dc circuitry of the instrument. In the meter instrument, the rectified current is averaged either by a filter or by the ballistic characteristics of the meter to produce a steady deflection of the meter pointer. Any dc component in the applied voltage is excluded from the measurement by an input-blocking capacitor preceding the high-gain amplifier.

The ac amplifier has a large amount of negative feedback, which ensures gain stability for measurement accuracy as well as broadening the frequency range of the instrument. Inclusion of the meter circuit in the feedback path minimizes the effects of diode nonlinearities and meter impedance variations on circuit performance.

It should be noted that the capacitors in the meter circuit tend to act as storage or filter capacitors for the rectifier diodes and also coupling capacitors for the feedback signal. However, both capacitors could be replaced by resistors and the circuit would work nearly as well with the inherent meter inertia for filtering. The diodes act as switches to maintain unidirectional meter current despite changes in the instantaneous polarity of the input voltage.

Errors in the readings of an average-responding voltmeter can be generally attributed to either the application of complex waveforms to the detector (a distorted or nonsinusoidal input) or the presence of nonharmonically related extraneous signals (hum and noise). It is not uncommon that a small amount of hum is combined with the voltage to be measured. When the frequency to be measured is relatively high with respect to the hum frequency, a small amount of hum, of about 10 percent, will increase the reading of the average-responding meter by about one-half as much as it would increase the reading on a true-rms meter. This

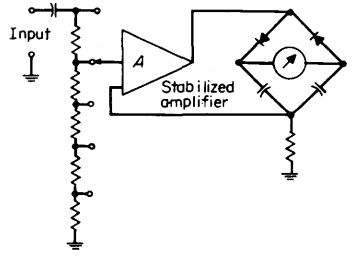


FIG 8-19 Block diagram of average-responding detector.

means a 10 percent hum will give a reading of approximately 1.0025 times the reading without hum, or an increase of only ¼ percent.

The voltage of thermal noise is characterized by a gaussian probability density function, $p(v)$. Shot noise also has a gaussian probability distribution if the average number of shots per second is much greater than the bandwidth. Impulse noise behaves as shot noise only if the impulses are totally independent and occur at random times. Rectified gaussian noise as obtained from an envelope detector, however, does not have a gaussian distribution.

We can calculate the indication of an average-responding voltmeter to gaussian noise as follows [1]: The rectified dc voltage V_0 from noise can be shown as

$$V_0 = \int_{-\infty}^{\infty} |v| p(v)\, dv \tag{8-7-1}$$

If $p(v)$ is symmetrical about zero, then

$$V_0 = 2\int_0^{\infty} vp(v)\, dv \tag{8-7-2}$$

If the noise is gaussian, the probability $p(v)\, dv$ that the instantaneous voltage lies between v and $v + dv$ is

$$p(v)\, dv = \frac{1}{\sigma(2\pi)^{1/2}} e^{-v^2/2\sigma^2} dv \tag{8-7-3}$$

where σ is the rms noise voltage. Substituting this in the expression for V_0 gives

$$V_0 = \frac{2}{\sigma(2\pi)^{1/2}} \int_0^{\infty} ve^{-V^2/2\sigma^2} dv = \frac{2\sigma}{(2\pi)^{1/2}} \tag{8-7-4}$$

Since the meter is calibrated to read the rms value of a sine wave, or

$$V_{\text{indicated}} = 1.11V_0 = \frac{1/(2)^{1/2}E_{\max}}{2/\pi E_{\max}} V_0 = \frac{\pi}{2(2)^{1/2}} V_0 \tag{8-7-5}$$

where V_0 is the average or rectified value. Then,

$$V_{\text{indicated}} = \frac{\pi}{2(2)^{1/2}} \frac{2\sigma}{(2\pi)^{1/2}} = \frac{\sigma(\pi)^{1/2}}{2} = 0.886\sigma \tag{8-7-6}$$

Thus, average-responding meters read aproximately 11 percent, or 1 dB, low on gaussian noise provided that no overload occurs on the peaks.

The accuracy with which an average-reading voltmeter will indicate the rms value of a wave with harmonic content depends not only on the amplitude of the harmonic but also on its phase and order. In the case of a fundamental wave and a single harmonic (generally second or third),

the waveform can be described as follows:

$$v = V_{max}[\sin \theta + k \sin (n\theta + \phi)] \tag{8-7-7}$$

where k is the amplitude of the harmonic as a percentage of the fundamental. We can solve for the average value by writing

$$V_{av} = \frac{V_{max}}{2\pi} \left\{ \int_{\theta_1}^{\theta_2} [\sin \theta + k \sin (n\theta + \phi)]\, d\theta - \int_{\theta_2}^{\theta_1 + 2\pi} [\sin \theta + k \sin (n\theta + \phi)]\, d\theta \right\} \tag{8-7-8}$$

which reduces to

$$V_{av} = \frac{V_{max}}{\pi} \left\{ \cos \theta_1 - \cos \theta_2 + \frac{k}{n} [\cos (n\theta_1 + \phi) - \cos (n\theta_2 + \phi)] \right\} \tag{8-7-9}$$

where θ_1 and θ_2 are zero crossing points determined by solving the following equation:

$$\sin \theta + k \sin (n\theta + \phi) = 0 \tag{8-7-10}$$

The reading that will appear on an average-responding voltmeter is obtained by multiplying Eq. (8-7-9) by the average-to-rms form factor 1.11. That is,

$$V_{rms} = \frac{1.11 V_{max}}{\pi} \left\{ \cos \theta_1 - \cos \theta_2 + \frac{k}{n} [\cos (n\theta_1 + \phi) - \cos (n\theta_2 + \phi)] \right\} \tag{8-7-11}$$

It should be noted that waveforms with large amounts of harmonic content may have more zero crossings than the fundamental. This is particularly true of third-harmonic distortion of greater than about 30 percent, second-harmonic distortion greater than 50 percent and higher-harmonic distortion. If this is the case, Eqs. (8-7-8), (8-7-9), and (8-7-11) must be modified to integrate the rectified waveform between zero crossings.

The error in rms reading as a function of the harmonic phase is shown in Fig. 8-20 for second-harmonic distortion and in Fig. 8-21 for third-harmonic distortion. Curves are shown for 10, 20, and 30 percent harmonic distortion in each of the two plots. It is interesting to note that cyclic variation occurs in the error as the phase of the harmonic is changed. (See Fig. 8-22 for phase orientation.) The maximum error with second-harmonic distortion occurs when the second harmonic is in phase or out of phase (180°) with the fundamental. The maximum error

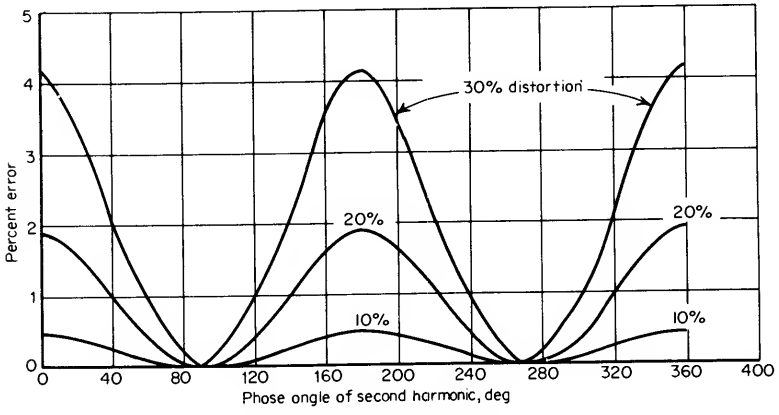


FIG 8-20 Error between true-rms and average reading × form factor of wave with second-harmonic distortion.

in a signal with third-harmonic distortion occurs when the phase of the harmonic is either in phase or out of phase (180°) with the fundamental. Figure 8-22 shows the effect on the waveforms caused by second- and third-harmonic distortion with the harmonic phase set to obtain the maximum error in the voltmeter readings.

The true-rms value of a waveform with harmonic distortion can be

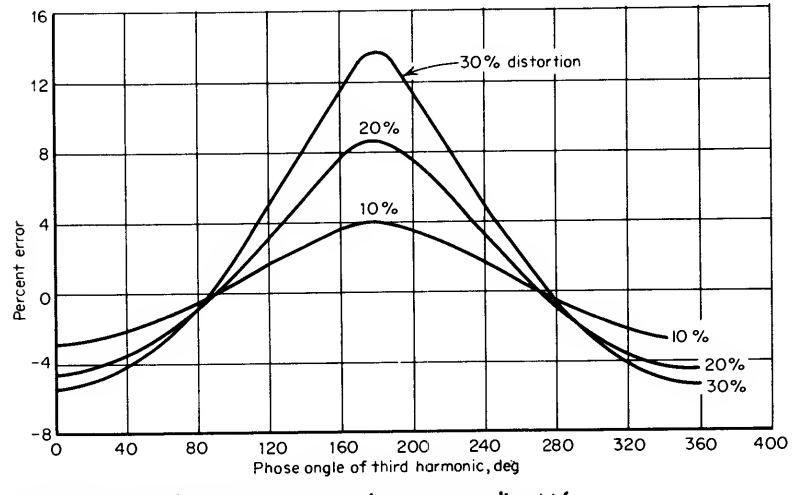


FIG 8-21 Error between true-rms and average reading × form factor of wave with third-harmonic distortion.

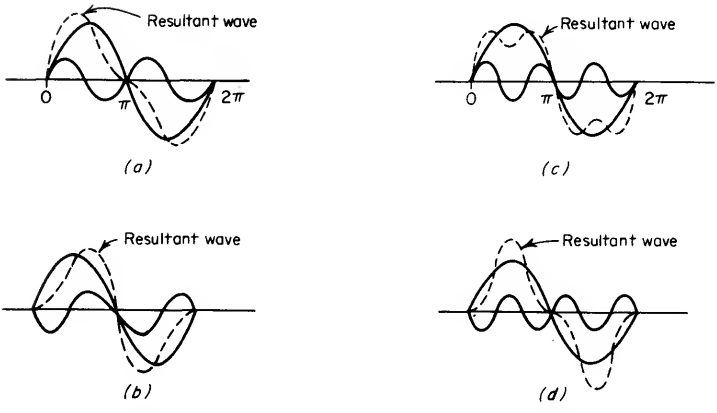


FIG 8-22 Distorted waveforms with largest time error in average-responding detector: (a) in-phase second harmonic, (b) out-of-phase second harmonic, (c) in-phase third harmonic, and (d) out-of-phase third harmonic.

calculated by taking the square root of the sum of the squares of the rms value of the fundamental and the rms values of the harmonics:

$$V_{\text{rms}} = [(V_{\text{rms}}, \text{fund})^2 + (V_{\text{rms}}, 2\text{nd})^2 + (V_{\text{rms}}, 3\text{rd})^2 + \cdots]^{1/2} \tag{8-7-12}$$

Table 8-1 shows the values read on a true-rms voltmeter of waveforms with second- and third-harmonic distortion and the maximum percentage error obtained on an average-responding voltmeter.

An examination of Figs. 8-20 and 8-21 and Table 8-1 indicates a number of interesting points. From Fig. 8-20 it can be seen that the condition in which a second harmonic will cause the average value of the complex wave to follow most closely the rms value is that the harmonic have a 90° relation to the fundamental, that is, where the peaks of the harmonic

TABLE 8-1

% Harmonic	True rms voltage	Maximum % error 2nd harmonic only	3rd harmonic only
1	0.7071	0.02	0.34
3	0.7074	0.032	1.03
5	0.7080	0.112	1.77
10	0.7106	0.485	3.8
20	0.7211	1.93	8.5
30	0.7382	4.2	13.8

occur at the time the fundamental intercepts the zero axis. This is also the condition often encountered in practice. A square-law term in a transfer characteristic, for example, produces this phase relation for the second harmonic. It can also be seen that for second harmonics of typical magnitudes (less than 10 percent), the average-reading meter will give readings quite close to the rms. For a second harmonic of 10 percent magnitude, for example, the error of the average-reading meter is less than 1 percent.

It can be noted from Fig. 8-21 that a wave consisting of a fundamental and third harmonic causes considerably greater variations in the readings of an average-reading type of voltmeter than does a wave with second-harmonic content. Whereas the reading of the meter on a wave containing second harmonic is always lower than the rms value, the reading with a wave containing third harmonic can be either high or low for harmonic contents up to as high as 75 percent.

Not only does the third harmonic cause greater variations in the meter reading than second harmonic, but it also causes greater variations than any other harmonic. The extremes of error with small amounts of odd harmonics are given by the percentage of the harmonic divided by the order of the harmonic. Small amounts of harmonic in this case can be defined as percentages less than $100/n$, where n is the order of the odd harmonic.

It should be noted that for typical amounts of this worst harmonic, the third, the accuracy of an average-reading meter is still good. Third harmonics up to 10 percent, for example, can cause errors no greater than 5 percent.

When more than one harmonic is present in the applied wave, the mathematics of each case becomes more complicated. As a result, no analytical studies of these situations have been made. However, some experimental data have been compiled for combined second and third harmonics with various amounts of fundamental [2].

Two other cases of interest that can be easily investigated are the differences between readings on an average-responding voltmeter and a true-rms voltmeter with a triangular wave or a square wave applied. A calculation of the average value of a triangular wave, shown in Fig. 8-23*a*, shows

$$V_{av} = \frac{2}{\pi}\int_0^{\pi/2} \frac{2t}{\pi}\,dt = \frac{1}{2} \tag{8-7-13}$$

The rms value of the triangular wave is

$$V_{rms} = \left(\frac{2}{\pi}\int_0^{\pi/2} \frac{4t^2}{\pi^2}\,dt\right)^{1/2} = \frac{1}{3^{1/2}} \tag{8-7-14}$$

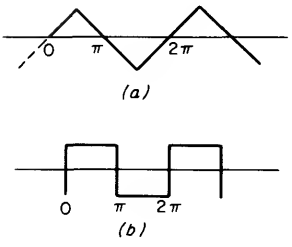


FIG 8-23 Triangular and square waveforms.

Multiplying the average value of the triangle wave by the form factor 1.11 and determining the error between the average-responding and true-rms voltmeters,

$$\text{Percent error} = \frac{1.11/2 - 1/(3)^{1/2}}{1/(3)^{1/2}} \times 100$$

$$= -3.81 \text{ percent} \quad (8\text{-}7\text{-}15)$$

In a square wave the unique relation exists that the average, rms, and peak values are all the same. Since an rms-calibrated, average-reading meter indicates 1.11 times the average value, it will indicate 11 percent high for the rms value of a square wave. Further, a square wave has the lowest ratio of rms value to absolute-average value of any wave. It follows, then, that an average-responding meter will never read more than 11 percent too high.

8-8 *Peak-responding Detectors*

The primary difference between the peak-responding voltmeter and the average-responding voltmeter is the use of a storage capacitor with the rectifying diode. The capacitor charges through the diode to the peak

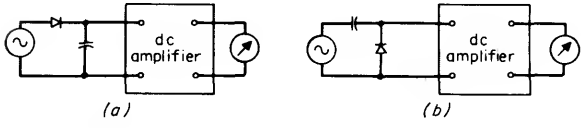


FIG 8-24 Peak-responding meters.

value of the applied voltage, and the meter circuit then responds to the capacitor voltage.

Two of the most common forms of the peak-responding detector are shown in Fig. 8-24. Figure 8-24*a* shows a dc-coupled peak detector, in which the capacitor charges to the total peak voltage above ground refer-

ence. In this case, the meter reading will be affected by the presence of dc with the ac voltage. In Fig. 8-24*b*, an ac-coupled peak detector circuit is shown. This circuit is closely related to the familiar dc-restorer circuits found in the literature.

In both circuits, the capacitor discharges very slowly through the high-impedance input of the dc amplifier so that a negligibly small amount of current supplied by the circuit under test keeps the capacitor charged to the ac peak voltage. The dc amplifier is used in the peak-responding meter to develop the necessary meter current.

The primary advantage of the peak-responding voltmeter is that the rectifying diode and storage capacitor may be taken out of the instrument and placed in the probe when no ac preamplification is required. The measured ac signal thus travels no farther than the diode. The peak-responding voltmeter, then, is able to measure frequencies up to hundreds of megahertz with a minimum of circuit loading. In fact, peak detectors with the circuit shown in Fig. 8-24*b* are being used in coaxial transmission-line configurations to measure signals with frequencies of greater than 40 GHz.

There are several inherent disadvantages of the peak-responding voltmeter for many applications. One is the susceptibility of the peak-responding meter to errors caused by harmonic distortion in the input waveform. Another is the limited sensitivity of the instrument because of the imperfect diode characteristics, which is discussed later. The third disadvantage appears when the input waveform is not symmetrical.

If the input waveform is unsymmetrical, a different reading will occur when the voltmeter leads are reversed. This is straightforward and obvious for the dc-coupled peak detector. One can envision a pulse train's being applied to the detector in which the peaks of a positive pulse would be indicated on the meter. However, what appears when a negative pulse train is ac coupled to the peak detector? When a positive pulse train is ac coupled to the detector, the "dc-restoring" property of the detector tends to restore the dc value lost in the coupling to the waveform. The output of the detector is then filtered to measure only the dc value of the detected waveform. If a pure sine wave or any other symmetrical waveform were measured, the detected and filtered value would be proportional to the peak-to-peak value of the waveform. For the positive pulse train, the filtered value would be approximately equal to the dc value of the original waveform. However, for the negative pulse train, the output of the detector prior to filtering would appear to be a positive pulse train with a duty factor approaching unity, as shown in Fig. 8-25. The filtered output is obviously close to the peak value of the input pulse train.

The limitation of sensitivity in the peak-responding detector is pri-

marily caused by the nonideal characteristics of the rectifying diode. Diodes, whether semiconductor or thermionic, have highly nonlinear current-to-voltage transfer characteristics below 1 V. This nonlinearity

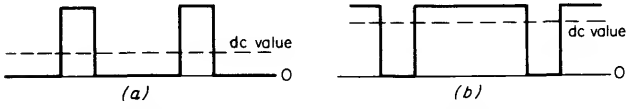


FIG 8-25 Alternating-current peak-detector output with pulse train applied to input: (a) positive train; (b) negative pulse train.

is sometimes compensated for by a separate nonlinear meter scale on the most sensitive range or by a differential technique with the nonlinearity of another closely matched diode. However, accuracy is difficult to achieve since individual diodes of a given type do not necessarily have similar transfer characteristics. In Fig. 8-26 assume that the capacitor C is sufficiently large that its voltage change during one cycle is negligible compared with the voltage change at the output. The voltage at the instantaneous output, v_o, consists of a dc value V_o equal to the dc voltage across the capacitor, V_d, and a series of ac terms related to the input voltage and its harmonics caused by the nonlinearity of the diode. We can write

$$v_o \approx V_d + V_a \sin \omega t \tag{8-8-1}$$

The dc component of the output is determined by the amount of charge added to the capacitor by the diode to replace that dissipated by the

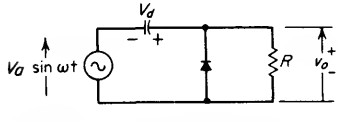


FIG 8-26 Alternating-current peak detector.

resistor during the positive half cycle. The instantaneous diode current is

$$\begin{aligned} i &= I_s(e^{qv_o/nkT} - 1) = I_s(e^{(q/nkT)(V_d+V_a \sin \omega t)} - 1) \\ &= I_s(e^{(q/nkT)(V_o+V_a \sin \omega t)} - 1) \end{aligned} \tag{8-8-2}$$

where n is a constant of proportionality determined by the diode manufacturing process. In many recently designed silicon diodes, n has been experimentally found to be approximately 2. In the equation, q is the charge on an electron, k is Boltzman's constant, and T is temperature in degrees Kelvin.

The average current through the diode, I, is equal to the current lost through the load resistor R during the positive half cycle; that is,

$$I = \frac{V_o}{R} = I_s(e^{(q/NkT)(V_o+V_a \sin \omega t)} - 1) \qquad \text{average} \tag{8-8-3}$$

or

$$\frac{V_o}{I_s R} = (e^{(qV_o/nkT)} e^{(qV_a/nkT) \sin \omega t} - 1) \qquad \text{average} \tag{8-8-4}$$

Rearranging,

$$\left(1 + \frac{V_o}{RI_s}\right) e^{qV_o/nkT} = e^{(qV_a/nkT) \sin \omega t} \qquad \text{average} \tag{8-8-5}$$

we can expand the right side of Eq. (8-8-5) in a Bessel function form

$$e^{jx \sin \theta} = J_0(x) + 2jJ_1(n) \sin \theta + 2J_2(x) \cos 2\theta + \cdots \tag{8-8-6}$$

or

$$e^{x \sin \theta} = J_0(-jx) + 2jJ_1(-jx) \sin \theta + \cdots \tag{8-8-7}$$

Then for the dc component, we can write

$$\left(1 + \frac{V_o}{RI_s}\right) e^{qV_o/nkT} = J_0\left(-j \frac{qV_a}{nkT}\right) = \frac{I_o q V_a}{nkT} \tag{8-8-8}$$

We can examine the peak detector under three separate conditions, depending on the level of the input signal. For high-level signals, one can intuitively predict a linear response to input voltage, since the diode voltage is limited in the forward bias condition. In most cases, where the resistance on the output of the detector is large (10 MΩ or greater), input voltages of greater than 1 V can be assumed to be linearly detected. If the input voltage is less than 1 V, the detected voltage can be determined to a close approximation by using Eq. (8-8-8). However, for input voltages less than 50 mV, the diode transfer characteristics can be considered square law by making the following approximations:

$$\frac{V_o}{RI_s} \ll 1$$

$$e^{qV_o/nkT} \approx 1 + \frac{qV_o}{nkT}$$

and

$$e^{(q/nkT)V_a \sin \omega t} = 1 + \frac{q}{nkT} V_a \sin \omega t + \frac{1}{2}\left(\frac{q}{nkT}\right)^2 V_a^2 \sin^2 \omega t + \cdots \tag{8-8-9}$$

Since we are interested only in the dc component, we can use Eq. (8-8-5) and write

$$1 + \frac{qV_o}{nkT} = 1 + \frac{1}{4}\left(\frac{q}{nkT}\right)^2 V_a^2 + \cdots$$

or

$$V_o = \frac{q}{4nkT} V_a^2 = \frac{q}{2nkT} V_{in}^2 \qquad \text{where} \qquad V_{in} = \frac{V_a}{2^{1/2}} \tag{8-8-10}$$

The accuracy of Eqs. (8-8-8) and (8-8-10) depends heavily on the manufacturing processes in the fabrication of the detector diode, but the equations are useful to determine the detected voltage to a reasonable approximation.

8-9 *Peak-to-peak Detection*

The addition of a capacitor and a diode can change a peak detector into a peak-to-peak detector, as shown in Fig. 8-27. The advantage of detecting peak to peak over a simple peak detection is the absence of turnover error caused by an unsymmetrical waveform. In addition, the output voltage for a symmetrical waveform is double the output voltage of a peak detector, which results in twice the detection efficiency.

The peak-to-peak detector is sometimes called a *voltage-doubler* circuit, and it works as follows: When the waveform at V_1 is negative, diode D_1 becomes forward biased and C_1 charges up to approximately the negative peak voltage. When V_1 goes positive, D_1 is back biased and D_2 becomes forward biased. The charge on C_1 is gradually transferred to C_2 during the initial transient period. When the circuit is in steady-state operation, the output voltage is the sum of the voltage developed across C_1 during the negative portion of V_1 and the positive peak of V_1, which is equal to the peak-to-peak input voltage. To ensure successful operation

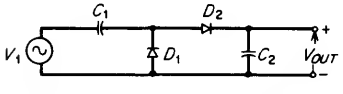


FIG 8-27 Peak-to-peak detector.

of the circuit, C_1 and C_2 must be large enough so that the voltage does not change appreciably across C_2 during one period of the input voltage, and the voltage across C_1 does not appreciably change in the process of recharging C_2.

8-10 Root-mean-square–responding Detectors

There are many occasions when measurements of the true rms value of a voltage are highly desirable. When measurements of electrical or acoustical noise, low duty-cycle pulse trains, or voltages of undetermined waveform are made, it is almost imperative that an rms-responding voltmeter be used.

In past years, rms-responding voltmeters have not been used widely because of the difficulties in designing a rugged, easily operated instrument of somewhat reasonable cost. Recently, however, these difficulties have been largely overcome.

The rms-responding voltmeter presents special circuit-design problems compared with the straightforward techniques used in the average- and peak-responding voltmeters. This is because the input voltage must be squared and then the square root of the average of the squared quantity taken.

One approach has been to take advantage of the nonlinear characteristics of diodes, which exhibit a fairly accurate square-law transfer function at low voltage levels. By calibrating the meter scale so that it indicates the square root of its driving voltage, one makes the meter indicate the rms value. Accurate calibration is difficult, however, because the diode characteristics do not always conform precisely to a square-law curve, and this characteristic is not uniform from diode to diode. This problem is reduced by amplifying the signal and using a greater voltage to drive the meter through a more predictable nonlinear network made up of several diodes and resistors.

Another approach is to use a thermocouple. The signal to be measured is applied to a fine heater wire, and a thermocouple attached to the heater wire generates a dc voltage proportional to the rise in temperature of the hot junction. This measurement is based on the original concept of the rms value as the equivalent of the heating power in a waveform.

The accuracy of this technique has been difficult to control because of the nonlinear behavior of thermocouples, which tends to complicate the meter calibration, and also because of the thermal problems involved. Thermal variations are reduced by installing the heater and thermocouple in an evacuated glass bulb and by using fine wires of low thermal conductivity. Other problems with thermocouples have been sluggish response and susceptibility to burnout.

One technique to reduce these difficulties is to use a null-balance technique as shown in Fig. 8-28. The amplified input signal is applied to the measuring thermocouple while a dc feedback current is fed to the heater of the balancing thermocouple. The dc current is derived from the voltage output difference between the thermocouples. The circuitry

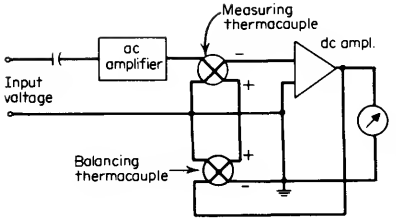


FIG 8-28 Block diagram of true-rms voltmeter with null-balance technique.

may be regarded as a feedback control system which matches the heating power of the dc feedback voltage to the input waveform heating power. Meter deflection is proportional to the dc feedback, which in turn is equivalent to the rms value of the input signal if the loop gain is high. The meter indication, therefore, is linear versus input and not subject to the nonlinearities of the thermocouples.

A frequent limitation on the usefulness of an rms-responding voltmeter for measuring highly nonlinear waveforms, such as pulse trains, is its crest-factor rating. Crest factor is defined as the ratio of the peak voltage to the rms voltage of a waveform. The crest factor is limited primarily by the amplifiers in the voltmeter circuit preceding the detector. The maximum crest factor is determined by (1) the level beyond which the input waveform drives the amplifiers into nonlinear operation and (2) the bandwidth of the amplifiers, which determines how much of the frequency spectrum of the pulse train or other nonlinear waveform will be accurately detected. A pulse train and its frequency spectrum are shown in Fig. 8-29. It should be observed that the width of the frequency spectrum is inversely proportional to the width of the pulse and, as will be shown below, is increased as the crest factor of the pulse train is increased. If the crest factor of the waveform is large and the repetition rate of the pulse train is within a decade of the bandwidth of the system, it is possible that an appreciable amount of the energy of the

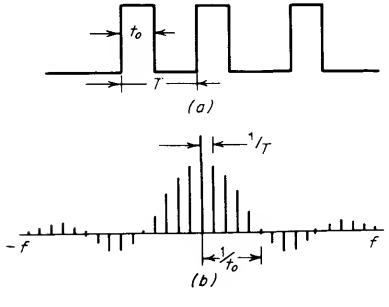


FIG 8-29 Frequency spectrum of a pulse train: (a) pulse train; (b) frequency spectrum.

input waveform would not pass through the amplifiers, which would result in an error in the voltmeter reading.

Since a pulse train represents an extreme case of a nonsinusoidal periodic waveform and in some cases is a reasonable approximation of impulse

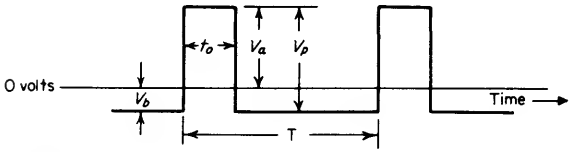


FIG 8-30 Pulse train.

noise, it can be considered essential in studying the effects of waveform crest factor on the accuracy of an rms-responding voltmeter. From Fig. 8-30 we can define duty cycle D as the pulse width t_0 divided by the waveform period T. That is,

$$D = \frac{t_0}{T} \tag{8-10-1}$$

Since most rms-responding voltmeters are ac coupled, we can assume

$$V_a + V_b = V_{pp} \tag{8-10-2}$$

and

$$V_a t_0 = V_b(T - t_0) \tag{8-10-3}$$

Then

$$V_b = V_{pp}D \tag{8-10-4}$$

$$V_a = V_{pp}(1 - D) \tag{8-10-5}$$

The rms sum of V_a and V_b is, after integration,

$$V_{\text{rms}} = \left(\frac{V_{pp}^2(1 - D)^2 t_0 + V_{pp}^2 D^2(T - t_0)}{T}\right)^{1/2}$$

$$= V_{pp}[D(1 - D)]^{1/2} \tag{8-10-6}$$

Since crest factor CF is equal to V_a/V_{rms} for values of $0 \leq D \leq 1/2$

$$CF = \frac{V_{pp}(1 - D)}{V_{pp}[D(1 - D)]^{1/2}} = \left(\frac{1}{D} - 1\right)^{1/2} \tag{8-10-7}$$

If the duty factor D is small compared with 1, the crest factor is approximately equal to the reciprocal of the square root of the duty factor, or

$$CF \approx \frac{1}{D^{1/2}} \tag{8-10-8}$$

Thus, the crest factor of a pulse waveform with low duty factor turns out to be somewhat different from that which might be presumed from consideration of the peak-to-average ratio. A pulse waveform with a duty factor of 1 percent has a crest factor of approximately 10, not 100 as might be assumed.

High crest-factor performance is not easily obtained. An rms voltmeter with a high crest-factor rating must have amplifiers with sufficient dynamic range to pass signals that have a peak amplitude many times larger than full-scale rms value. For example, a pulse train with a crest factor of 7 must have a minimum dynamic range corresponding to 7.14 V to read full scale on the 1-V range of an rms voltmeter and a dynamic range corresponding to 14 V to read correctly if the pulse train should be turned over.

8-11 Other Detection Methods

There are several detecting schemes used to obtain so-called true-rms readings by means other than measuring the heating value of a waveform with a thermocouple or other heat-sensing device. The most common technique uses a combination of the average detector and peak detector with suitable proportionality factors determined by the waveforms most frequently measured. This is shown in Fig. 8-31.

For any given waveform, the rms value can be expressed by its peak and average values. That is,

$$V_{\text{rms}} = k_1 V_p + k_2 V_a \tag{8-11-1}$$

where k_1 and k_2 are proportionality factors for the peak and average values of the input waveform, respectively. Dividing by V_a,

$$\frac{V_{\text{rms}}}{V_a} = k_1 \frac{V_p}{V_a} + k_2 \tag{8-11-2}$$

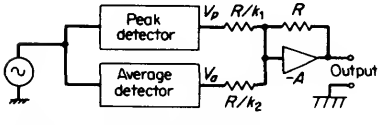


FIG 8-31 Quasi-rms detector.

Using Eq. (8-11-2), we can solve for k_1 and k_2 for any two waveforms, knowing their form factor and peak-to-average ratio. It should be noted that the detector will indicate the true rms value for only the waveforms whose form factor and peak-to-average ratio are used in the design. Any

other type of waveform will be detected with some error. If the waveforms chosen are representative of those to be measured, the detector will read the rms value with a relatively small amount of error, which can be calculated by using Eq. (8-11-1).

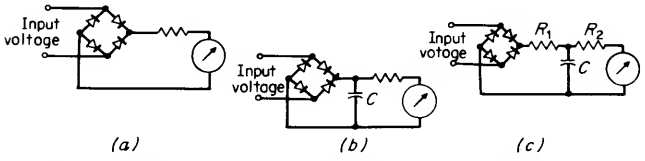


FIG 8-32 Combined peak and average detector with rms output: (a) average detector, (b) peak detector, and (c) combined detector.

This type of detector is most commonly used for special applications, such as measuring the rms value of a two-tone signal in telephone work or measuring random noise.

Another technique closely related to the scheme discussed above is one in which the average and peak detectors are combined as shown in Fig. 8-32. Figure 8-32*a* shows a simplified average detector and Fig. 8-32*b* shows a peak detector. Figure 8-32*c* shows the combination of the average and peak detectors in which the approximation to the rms value is determined. The ratio R_1/R_2 can be adjusted empirically to provide the rms output for a group of closely related waveforms over a limited dynamic voltage range and a limited frequency range.

Another possibility is to make R_1 nonlinear by using a diode shaping network or other suitable nonlinear network to more closely approximate a true rms reading over a large dynamic range. The shaping network could be used to "square" the input voltage, which is averaged by the average detector, and applied to a meter with scale resembling a square-

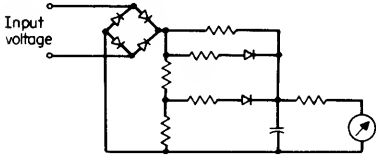


FIG 8-33 Quasi-rms detector with nonlinear R_1 substituted in previous figure.

root function. This would closely approximate the root-mean-square process over a relatively large dynamic range, depending on how closely the shaping network produces the desired nonlinear effect. This circuit, shown in Fig. 8-33, is also primarily designed by empirical means.

8-12 Sampling Voltmeters

Sampling techniques used to construct low-frequency equivalents of high-frequency waveforms have been in existence for several years in oscilloscopes to display the waveforms of very high frequency repetitive signals. Only within the past few years has this technique been applied to voltmeters. An rf vector voltmeter is available that uses a sampling technique to measure amplitudes and phase angles simultaneously and automatically at frequencies as high as 1 GHz. Other sampling instruments are being developed to operate at frequencies as high as 12.4 GHz.

Most sampling instruments, including the sampling oscilloscope and the rf vector voltmeter, sample coherently. This is analogous to the familiar stroboscopic technique, by which an oscillating or repetitive motion is apparently "slowed down" by observing it only at discrete times, instead of continuously. The observations, or samples, may be taken by flashing a light, by observing the oscillating object through a slit in a rotating disc, or by some other means. Coherent sampling is shown in Fig. 8-34*a* and *b*. This type of sampling is primarily used when the waveform must be preserved for visual presentation or phase measurement. In a sampling voltmeter designed to measure only magnitude, it is advantageous to sample incoherently, because the voltmeter can be made very broadband without requiring frequency tuning. The input voltage is sampled at irregular intervals that have no relationship to any of the frequency components of the input signal. Enough samples are taken, however, so that the average, peak, and rms values of the samples closely approximate the average, peak, and rms values of the input voltage. Thus the information that is relevant to the voltage-measuring function is preserved, while the waveform, which is assumed irrelevant, is not.

Incoherent sampling, shown in Fig. 8-34*c* and *d*, is especially advantageous in a voltmeter, because it gives the meter the sensitivity, accuracy, and broad frequency range of a sampling instrument, and yet it is less costly than coherent techniques and, unlike coherent sampling, it does not require that the input signal be periodic. The sampling voltmeter operates equally well with sinusoidal, pulsed, random, or FM signals.

For the technique of incoherent sampling to work in all situations it is necessary that there be no correlation between the sampling times and the motion or signal under observation. If the sampling frequency were a subharmonic of the frequency of the signal being measured, the motion would be completely stopped. Thus, all the samples would be exactly the same height, and it would be impossible to determine the peak, average, rms, and so on. One way to produce uncorrelated or nonuniform sampling intervals is to frequency-modulate the basic sampling signal

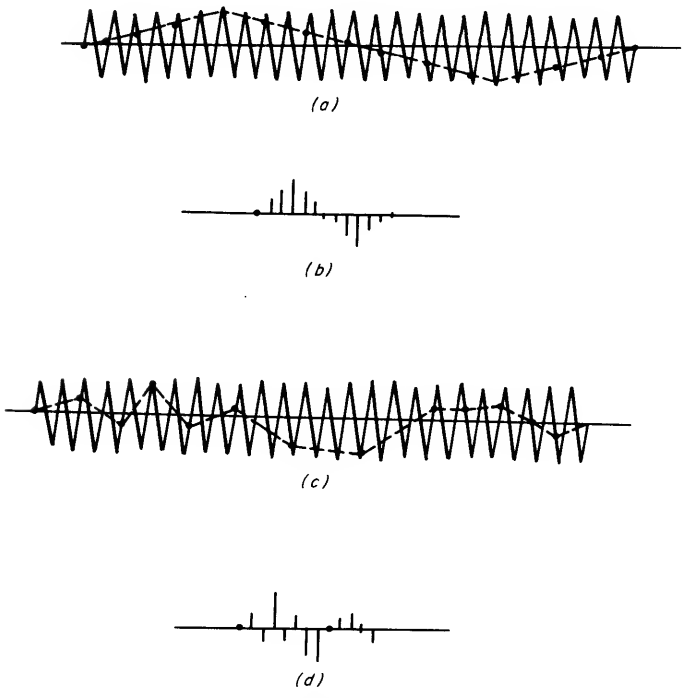


FIG 8-34 Sampled outputs with coherent and random sampling: (a) input waveform, coherent sampling; (b) sampled output, coherent sampling; (c) input waveform, random sampling; (d) sampled output, random sampling.

with a low-frequency triangular wave. This produces sampling intervals which are for all practical purposes uncorrelated with all input signals.

A block diagram of a typical sampling voltmeter is shown in Fig. 8-35. Incoherent intervals are generated by frequency modulating the sampling rate at a 10-Hz rate. The sample hold circuit retains a constant voltage proportional to the sample until the next sampling instant.

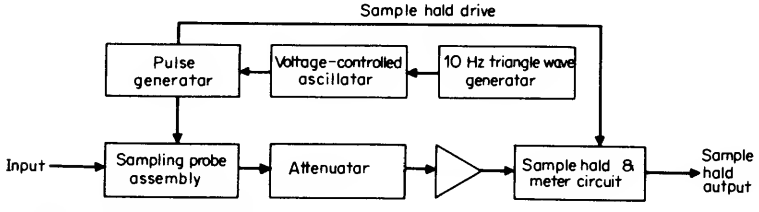


FIG 8-35 Sampling voltmeter.

8-13 Synchronous Detection

A difficult but essential voltage measurement required for many applications is that of measuring low-level signals obscured by noise or other nonrelated signals. Such conditions are frequently encountered in communications systems, medical research, and control systems. Unfor-

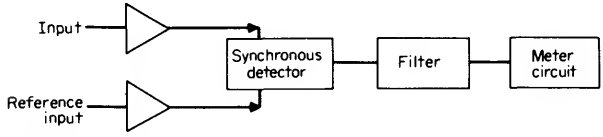


FIG 8-36 Synchronous rectifier.

tunately, almost all broadband voltmeters are limited in sensitivity by noise and spurious signals introduced along with the input signal, caused by interference and ground loops within the system or by noise inherent in the amplifying circuitry.

One common approach to this type of measurement is the use of a synchronous rectifier driven by a reference signal at the same frequency as the fundamental or most significant frequency in the input signal, as shown in Fig. 8-36. Voltmeters using this technique require a clean, high-level reference-signal input from the test-signal source, or from a local oscillator inside the voltmeter. This type of voltmeter is generally average responding and calibrated in rms volts, but it is conceivable that true-rms-responding synchronous detecting voltmeters could be easily designed.

Another approach, somewhat unique to this type of measurement, is the phase-locked synchronous detector, in which no external reference signal is required. An internal voltage-controlled oscillator is used to drive the synchronous detector, as shown in Fig. 8-37. The internal

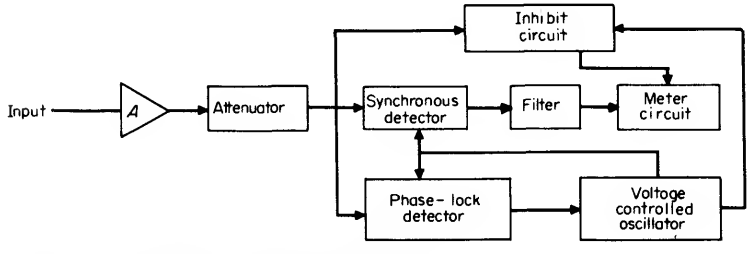


FIG 8-37 Block diagram of phase-lock voltmeter.

oscillator is coarsely tuned to the frequency of interest in the input wave, where the phase-lock loop automatically locks onto the input signal. The voltage controlling the frequency of the internal oscillator is derived from a phase detector, in which the input signal and internal oscillator output are compared in phase.

Ordinarily, a synchronous detector (or demodulator) is simply a phase-sensitive detector in which the reference signal is in phase with the signal to be demodulated. This relationship gives maximum efficiency of amplitude demodulation; to get maximum efficiency of *phase* detection, the reference signal should be in quadrature with the input signal. Therefore, in Fig. 8-37 it can be assumed that the phase-lock detector contains a 90° phase shifter.

The synchronous detector can in most cases be considered a product modulator. That is, the detector output voltage is proportional to the product of the input signal and the reference signal. If the detector is tuned to the frequency of interest in the input signal, this component of the input is frequency transformed to dc, filtered by a low-pass filter, and read out on the meter face. Any frequency not harmonically related to the frequency of the reference signal is averaged to zero and will not be read. Hence, noise and spurious signals are rejected.

Usually, the synchronous detector is driven by a square-wave reference signal, either shaped by an internal amplifier or obtained directly from the internal oscillator. This is done to maximize the detection efficiency of the circuit. To understand the response of the detector to a general input signal, an examination of this technique must be made. The reference signal can be expressed as

$$v_{\text{ref}} = V_{r_1} \sin \omega_1 t + V_{r_3} \sin 3\omega_1 t + V_{r_5} \sin 5\omega_1 t + \cdots \qquad (8\text{-}13\text{-}1)$$

and a general input signal as

$$v_i = V_1 \sin \omega_1 t + V_2 \sin 2\omega_2 t + V_3 \sin 3\omega_2 t + \cdots \qquad (8\text{-}13\text{-}2)$$

It can be seen that upon multiplying Eqs. (8-13-1) and (8-13-2) and then taking the average, several things become apparent:

1. No dc due to even harmonics of the input signal appears in the detected output. This is true because even harmonics have integral numbers of cycles in one period of the fundamental and the average voltage is zero.

2. Only odd harmonics of the input signal will contribute dc terms in the detected output, the amplitudes of which are inversely proportional to their harmonic number.

3. If the reference oscillator is tuned to a lower frequency than the frequency of interest in the input signal, so that the frequency of interest is odd-harmonically related to the reference oscillator frequency, a meter

indication will appear even though there is no energy in the input signal at this frequency. This is an undesired effect and is generally suppressed by an inhibiting circuit.

The synchronous detection, or frequency-selective voltmeter, is an extremely valuable instrument to use for measuring signals in the presence of noise and spurious signals. However, care must be taken in the interpretation of the meter reading, particularly when the input voltage is not sinusoidal.

8-14 Direct-current Probes

It is frequently undesirable to connect a known resistance in series with a conductor in order to measure the dc component (time average) of current by measuring the voltage across that resistor, and a moving-coil indicating instrument can be just as undesirable or inconvenient. The alternative is to measure the average magnetic field[1] produced by the current in the vicinity of the conductor, and this is not easy to do accurately.

Unless a magnetic path of low reluctance is provided around the conductor, the field measurement varies greatly as the distance between the conductor and the magnetic sensor changes. Of course, a torroidal core surrounding a conductor will concentrate the flux, but ordinarily one would have to break the circuit to insert the core or leave the core permanently installed. Some years back, however, a tiny split core was designed into a probe (Fig. 8-38) in such a manner that the two halves can be opened like jaws by squeezing the flanges on the probe handle (Hewlett-Packard model 428A). A spring return closes the jaws tightly around a conductor when the flanges are released. Practically no air gap remains.

The two halves of the core contain windings that are connected as a magnetic amplifier [3]. That is, an ac excitation is applied to the wind-

[1] More precisely, it is a line integral which is being evaluated, $\oint H dl = I$.

FIG 8-38 Probe jaws are opened by flanges on probe body; spring return closes jaws when flanges are released.

ings, but they are balanced and no output voltage occurs if the square-loop magnetic materials in the two halves saturate at the same electrical angles of the excitation voltage. However, when a dc magnetic field is produced by the current in the surrounded conductor, one-half of the core saturates earlier and the other half saturates later. An output voltage is produced by the magnetic amplifier at a frequency twice that of the excitation. A flexible cable transmits this output signal back to an instrument where it is amplified, demodulated, and then fed to an indicating meter and output terminals, where the output is available for an oscilloscope or a recorder [4]. Since the carrier frequency is about 40 kHz, the demodulated signal easily extends from dc to about 400 Hz. Bandwidth is limited to this value to keep noise low.

The dc probe does not introduce any dc resistance into the circuit under test. There are, however, some small secondary effects. The commercial unit already cited introduces an inductance of about 0.5 μH, a shunt capacitance to ground of about 2 pF, and an induced carrier-related voltage of up to 15-mV peak. These effects are rarely great enough to impair the accuracy (within ± 3 percent typically).

Typical full-scale current ranges are from 10^{-3} to 10 A. The sensitivity can be increased by looping several turns of wire through the probe, as shown in Fig. 8-39, although this arrangement is of course not always convenient. One of the basic limitations on sensitivity is the inherent noise in the magnetic amplifier. Another limitation is the effect of the earth's magnetic field. Good magnetic shielding and a symmetrical arrangement of the coils help to minimize this effect, but there is a practical limit. When the jaws of the probe are held open, the earth's field is equivalent to about 1 A, and this influence has been reduced to about 10^{-5} A by the procedures given above.

FIG 8-39 Measurement sensitivity can be increased by looping the conductor through the probe one or more times to increase the effective magnetizing force of measured current.

It is also possible to measure small magnetic fields by the Hall effect in semiconductors, but this approach has not been used in dc measurement as much as the magnetic amplifier has been used.

8-15 *Alternating-current Probe*

Whereas the dc probe utilizes a magnetic amplifier and has a frequency response limited to a fraction of the carrier frequency, the ac probe is

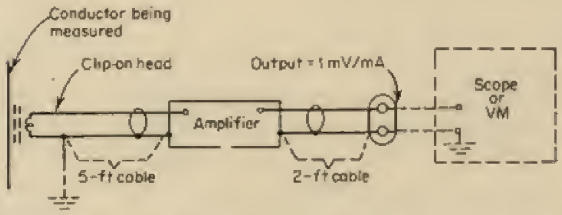


FIG 8-40 Current probe consisting of clip-on head and transistorized amplifier which can be used with oscilloscope or ac voltmeter.

essentially a broadband current transformer with a one-turn primary winding. An amplifier accompanies the probe to produce a proportional ac voltage (Figs. 8-40 and 8-41). This voltage can then be measured

FIG 8-41 Close-up view of ac probe (Hewlett-Packard 456A).

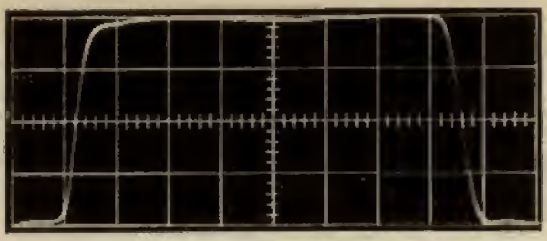


FIG 8-42 Fast pulse response of current probe as displayed by wideband oscilloscope. Sweep time is 0.05 μsec/cm, showing probe 10-to-90 percent rise time of 0.02 μsec.

with an oscilloscope or a voltmeter. Figure 8-42 shows an oscillogram made while using a current probe. One commercial unit has a scale factor of 1 mV/mA [5]. The frequency response has the normal bandpass characteristic to be expected of a transformer. This is illustrated in Fig. 8-43.

The loading effect of the commercial unit cited is less than $50 \times 10^{-3}\ \Omega$ and 0.05 μH in series with the measured wire. In addition, there is a shunt capacitance of approximately 4 pF to ground.

Broadband noise is less than 50-μA rms. Better sensitivity can be achieved by looping additional turns through the probe head. This raises the loading effect of the probe by raising the primary inductance. This increased inductance in combination with loop-to-loop capacitance can result in a resonance in the low-megahertz region, which limits the usefulness of adding more turns. An alternate way to circumvent the noise level limitation for sine-wave measurements is to use the current probe in conjunction with a wave analyzer. Some wave analyzers generate a signal identical in frequency with the input frequency setting. Use of this signal to stimulate the device under test ensures that the narrow bandwidth of the analyzer can be used to improve the signal-to-noise ratio. Measurement of currents as low as a few microamperes can be

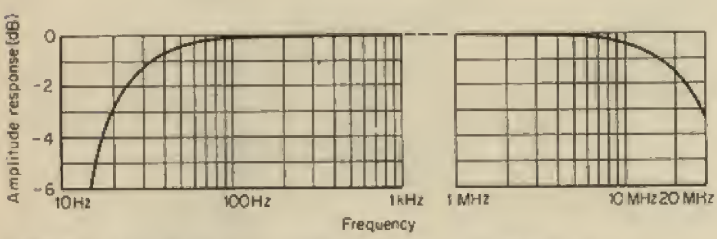


FIG 8-43 Typical frequency response of current probe when operated into rated load.

made in this way. The full bandwidth capability of the probe can still be used.

The probe can also be used for current summing and equalizing. Since the probe is effectively a current transformer, it has the property that it will sum the instantaneous values of the currents in two or more conductors that it may be clipped around. This property makes the probe a valuable and easily applied tool in applications in which it is desired to equalize alternating currents.

CITED REFERENCES

1. Hewlett-Packard Application Note 60, Which AC Voltmeter, January, 1965.
2. Oliver, B. M.: Some Effects of Waveform on VTVM Readings, Hewlett-Packard Application Note 60, app. 2, January, 1965.
3. Bergh, Arndt, et al: A Clip-on DC Milliammeter for Measuring Tube and Transistor Circuit Currents, *Hewlett-Packard J.*, vol. 9, June–July, 1958.
4. Barkley, D. E., and Arndt Bergh: Broader Information Capabilities in the Clip-on DC Milliammeter, *Hewlett-Packard J.*, vol. 13, November–December, 1961.
5. Forge, Charles O.: A New Clip-on Oscilloscope/Voltmeter Probe for 25 Hz to 20-MHz Current Measurements, *Hewlett-Paekard J.*, vol. 11, July–August, 1960.

CHAPTER NINE

IMPEDANCE MEASUREMENT

Henry P. Hall

General Radio Company, Concord, Massachusetts

Throughout the frequency spectrum, the measurement of impedance or its reciprocal, admittance, is as important as any other electrical measurement. Complex ratios of voltage to current characterize not only one-port devices or terminals of networks, but also multiport devices and systems. A rather thorough understanding of impedance measurement is essential to the technical man who must work with quantitative values for electrical networks and apparatus.

The chapter may seem too elementary to advanced students in the early sections, but it was considered desirable to present a comprehensive treatment. For the reader with a sound knowledge of impedance formulas and impedance components and simple meter measurements, the material on bridges and other sophisticated instruments is self-consistent and may be read alone.

9-1 Definitions and Formulas

Complex Impedance and Admittance. At dc, the resistance of a linear two-terminal device is defined as the ratio of the voltage across it to the

current through it by Ohm's law $R = E/I$. This quantity is often called the dc *resistance* R_{dc}. For sinusoidal ac, the ratio of voltage to current in general is complex. The ac equivalent of Ohm's law in cartesian form is $E/I = Z = R + jX$, where Z is called the *impedance* of the device. The real, or dissipative, part of impedance is referred to as the *effective*, or *ac*, *resistance* (sometimes written R_e or R_{ac}). The imaginary part is called *reactance* and represents the energy-storage part of impedance. These quantities R and X are both functions of frequency. At dc, X is either zero or infinite.

The reciprocal of dc resistance is dc *conductance*, and the reciprocal of impedance is called *admittance* Y. The latter is complex and has a real part called *effective*, or *ac*, *conductance* G, and an imaginary part called *susceptance* B. However, G and B are not reciprocals of R and X respectively, for

$$Y = G + jB = \frac{1}{Z} = \frac{1}{R + jX} = \frac{R}{R^2 + X^2} - j\frac{X}{R^2 + X^2} \tag{9-1-1}$$

Because all these quantities are frequency dependent, a stated value for any one of them is not a complete specification unless the frequency is given.

Polar Form. Impedance can also be expressed in polar as well as cartesian form; the relationships between them are

$$Z = R + jX = |Z|e^{j\theta} = |Z|(\cos\theta + j\sin\theta) \tag{9-1-2}$$

where the impedance magnitude $|Z| = \sqrt{R^2 + X^2}$, and the impedance phase angle $\theta = \tan^{-1} X/R$.

Likewise,

$$Y = G + jB = |Y|e^{j\phi} = |Y|(\cos\phi + j\sin\phi) \tag{9-1-3}$$

where

$$|Y| = \frac{1}{|Z|} = \sqrt{G^2 + B^2} \qquad \text{and} \qquad \phi = -\theta = \tan^{-1}\frac{B}{G}$$

We shall define the storage factor Q and the dissipation factor D by

$$Q = \frac{1}{D} = \frac{X}{R} = \frac{B}{G} \tag{9-1-4}$$

Factor D is always given as a positive number (for passive components), but the Q of a capacitive resistor is occasionally said to be negative. Of course, power factor is $D/(1 + D^2)^{1/2} = \cos\theta$.

Ideal Elements and Actual Components. Actual physical components

(devices) such as resistors, capacitors, and inductors are each made up of all three impedance parameters: resistance, capacitance, and inductance. These parameters are abstract mathematical quantities and never actually exist alone in pure form. However, it is very useful to hypothesize *ideal* or *pure* resistors, capacitors, and inductors, which each exhibit only the properties of the one respective parameter. These ideal quantities are called *resistance, capacitance, and inductance elements.* The symbols , , and are used to represent both the abstract elements and the actual components.

Equivalent Circuits—Series and Parallel. At any one frequency, a complex impedance can be described by two quantities, one real and one imaginary and, therefore, may be simulated by two ideal circuit elements, an *equivalent*[1] resistance and an *equivalent* capacitance or inductance. If each element represents one term in the expression $R + jX$, then they are assumed to be connected in series, for impedances of series elements are additive.

This impedance may also be expressed as an admittance by means of Eq. (9-1-1). If this admittance is represented by two elements, one representing G and the other B, then these elements are assumed to be connected in parallel.

This parallel circuit has the same total impedance as the series circuit and uses the same types of components, but the component values are different. That is, the equivalent series capacitance is not equal to the equivalent parallel capacitance, nor is the equivalent series resistance equal to the reciprocal of the equivalent parallel conductance (which is called the *equivalent parallel resistance*). The relationships between the series and parallel quantities (which could all be derived from Eq. (9-1-1)) are given in Table 9-1.

An example may be useful to emphasize this important point. If a

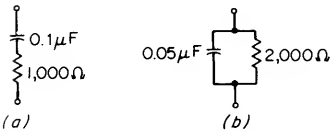


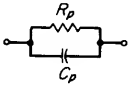
FIG 9-1 Series and parallel circuits that are equivalent at 1,592 Hz.

"black box" has a measured impedance of $1{,}000 - j1{,}000\ \Omega$ at 1,592 Hz, it acts at that frequency as the series connection of the ideal components of Fig. 9-1*a*. This impedance is equivalent to an admittance of 500 +

[1] The words *effective* and *equivalent* are often used interchangeably.

TABLE 9-1 Series-Parallel Conversion Formulas

Capacitance and resistance

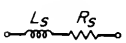


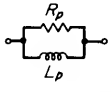
$$C_s = (1 + D^2)C_p \qquad C_p = \frac{1}{1 + D^2} C_s$$

$$R_s = \frac{D^2}{1 + D^2} R_p = \frac{1}{1 + Q^2} R_p \qquad R_p = \frac{1 + D^2}{D^2} R_s = (1 + Q^2)R_s$$

$$D = \frac{1}{Q} = \omega R_s C_s = \frac{1}{\omega C_p R_p}$$

Inductance and resistance





$$L_s = \frac{Q^2}{1 + Q^2} L_p = \frac{1}{1 + D^2} L_p \qquad L_p = \frac{1 + Q^2}{Q^2} L_s = (1 + D^2)L_s$$

$$R_s = \frac{1}{1 + Q^2} R_p \qquad R_p = (1 + Q^2)R_s$$

$$Q = \frac{1}{D} = \frac{\omega L_s}{R_s} = \frac{R_p}{\omega L_p}$$

$j500\ \mu\mho$, which can be represented by the parallel circuit of Fig. 9-1*b*. These two networks are indistinguishable at that frequency, but their component values are quite different.

These two-element equivalent circuits only represent the impedance of a physical element exactly at one frequency. One of the circuits may be reasonably accurate over a reasonable frequency range, but both cannot be. The more useful one is the one which better simulates the actual impedance.

Two measurements on the same impedance at two different frequencies will give four quantities, two real and two imaginary. Therefore, two ideal resistors and two ideal reactive elements can represent the impedance exactly at these two frequencies. These four elements may be connected in many different ways, and the best arrangement should be useful over a wide frequency range. Likewise, $2n$ elements are required to simulate exactly an impedance at n frequencies, but the diagram quickly becomes cumbersome. Equivalent circuits are given for the R, L, and C

components in Sec. 9-2, which use several elements, each representing a physical cause of impedance variation.

Transfer Impedance of Multiterminal Networks. The definition of impedance can be extended to include the ratio of any voltage to any current of the same frequency even if they do not both appear at the same terminals. If a current applied to one pair of terminals causes an open-circuit voltage to appear at another pair, the ratio of response voltage to applied current is called here the *open-circuit transfer impedance* Z_{to}. Likewise, if a voltage is applied and a short-circuit current measured, the ratio of response current to applied voltage is the short-circuit transfer admittance Y_{ts}. The open-circuit transfer admittance Y_{to} and the short-circuit transfer impedance Z_{ts} are the reciprocals of these two quantities respectively, but note that Y_{to} is not equal to Y_{ts} nor is Z_{to} equal to Z_{ts}.

While there are instruments that will measure these general transfer impedances, some specific cases are particularly important:

1. *Mutual Inductance M.* This is the open-circuit transfer inductance between two coupled coils. Actually, in coupled coils, the transfer impedance never represents a pure inductance, particularly for iron-core transformers, so that both series and parallel representations are possible (see Fig. 9-24).

2. *Transconductance* g_m. This term is often used to represent the real part of Y_{ts} of a vacuum tube or transistor.

3. *Four-terminal Impedance.* The impedance Z in Fig. 9-2 could be accurately measured, without error caused by impedances Z_1, Z_2, Z_3, and Z_4, by an instrument that could measure the open-circuit transfer impedance of this network. Such a network nicely "defines" a low-valued impedance Z because anyone making a proper transfer impedance

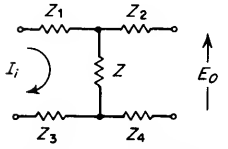


FIG 9-2 A four-terminal impedance $Z = E_o/I_i$.

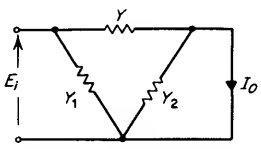
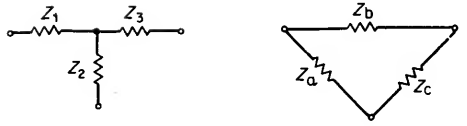


FIG 9-3 A three-terminal admittance $Y = I_o/E_i$.

or *four-terminal* measurement will get the same value no matter what impedance they add in series with the terminals.

4. *Three-terminal Admittance.* The admittance Y in Fig. 9-3 could be accurately measured without error from Y_1 or Y_2 by an instrument that could measure the short-circuit transfer admittance of this network.



$$Z_2 = \frac{Z_a Z_c}{Z_a + Z_b + Z_c} \qquad Z_b = Z_1 + Z_3 + \frac{Z_1 Z_3}{Z_2}$$

$$Y_2 = \frac{1}{Z_2} = Y_a + Y_c + \frac{Y_a Y_c}{Y_b} \qquad Y_b = \frac{1}{Z_b} = \frac{Y_1 Y_3}{Y_1 + Y_2 + Y_3}$$

(a) (b)

FIG 9-4 The $Y\Delta$ transformation.

This network nicely defines a low-valued admittance because stray capacitance or conductance between connections at either terminal does not affect the transfer admittance or *three-terminal* measurement.

YΔ Transformation. A Y (or T) network is equivalent to a Δ (or π) network if the branch impedances of the two networks are related by the formulas given in Fig. 9-4. The relationships are given for only one branch of each network because the others may be easily found by symmetry.

This transformation is very useful in reducing many networks, particularly complicated bridges, to a more understandable form. Bridge designers make use of it to simulate impedances that are difficult or impossible to obtain.

9-2 Components and Standards

An understanding of the behavior of what is being measured helps in the recognition of erroneous conclusions, for results that cannot be explained should be examined more closely. The purpose of this section is to give a brief description of how the value of components may vary under different conditions. It is not intended to explain the causes of these variations, which may be found in the references given.

9-2-1 *Resistors* [1, 2]

Behavior of Resistors. Resistors will change value as a result of applied voltage, power, ambient temperature, frequency change, or mechanical shock or humidity.

The *voltage coefficient* is the rate of change of resistance due to applied voltage, given in percent per volt. This characteristic is negative for most resistors, although some semiconductor devices actually increase in

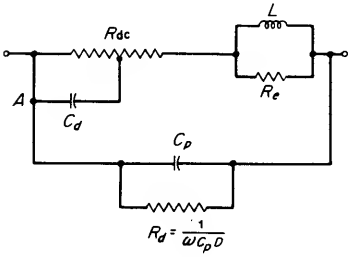


FIG 9-5 An equivalent circuit of a resistor.

resistance with applied voltage. The voltage coefficient of very high valued carbon-film resistors is usually rather large, while in wire-wound types the effect is usually negligible. Varistors are resistive devices designed to have a large voltage coefficient.

The *temperature coefficient* is the rate of change of resistance with ambient temperature, given in percent or parts per million (ppm) per degree Celsius. Many types of resistors increase in value as temperature is increased, while others, particularly hot-molded carbon types, have a maximum or minimum in their resistance curves which gives a zero temperature coefficient at some temperature. Metal-film and wire-wound types generally have temperature-coefficient values of less than 100 ppm/°C. Thermistors are resistance devices designed to have a large temperature coefficient.

The *power coefficient* is the product of temperature coefficient and temperature rise per watt, which gives a power coefficient in percent per watt and indicates the change in value resulting from applied power.

Frequency Characteristics of Resistors. Resistors change value with frequency because of inductance, lumped and distributed capacitance, dielectric loss, skin effect, and eddy-current losses, plus a few other minor effects as well [1, 3]. A rather complicated circuit diagram is given in Fig. 9-5 which accounts for the first-order difference terms resulting from these effects. Such a circuit would rarely be used, for not all the causes of change would be noticeable at any given resistance level.

Inductance L in series with a resistor does not change the equivalent *series* resistance but does change the equivalent *parallel* resistance, $R_p = R_{dc}\ (1 + Q^2)$, where $Q = \omega L/R_{dc}$. This effect is most noticeable in low-valued resistors, particularly wire-wound types that are not bifilar or Ayrton-Perry wound [4].

Lumped parallel capacitance C_p does not change the equivalent parallel resistance but does vary the equivalent series value,

$$R_s = \frac{R_{dc}}{1 + \omega^2 R_{dc}^2 C_p^2}$$

This effect is particularly important at high resistances. At medium values of resistance it is possible to have resonance between L and C_p which can cause an apparent increase in the effective series resistance. However, unless the resistor has particularly large inductance and capacitance, this effect is masked by others.

Distributed capacitance in a resistor is represented by a lumped capacitance C_d which can simulate the first term in a power series in $(j\omega)^2$ resulting from capacitance between various parts of the resistor. This capacitance reduces both series and parallel values of resistance. If the distributed capacitance were between the resistor and a third terminal, such as a shield (point A disconnected from the resistor terminal and becoming a third terminal), the transfer (direct) impedance has an increased effective inductance resulting from C_d, which is explained by the YΔ transformation (see Sec. 9-1).

Dielectric loss in both C_p and C_d can explain a change in resistance varying with the first power of frequency, $R = R_{dc}/(1 + \omega CDR_{dc})$ if the dissipation factor D is assumed constant. Measured changes in rf resistance usually show a slope somewhere between the first and second power of frequency, which suggests that capacitance and loss effects occur simultaneously [1].

Eddy-current losses of several types are represented by R_e, which is shown as loss in the series inductance. This is an accurate physical description of eddy-current losses in nearby conductors but *skin effect*, caused by inductance inside the conductors of the resistor itself, is usually explained by a somewhat different mechanism [3, 5]. However, R_e acts as the first-order term of the skin-effect change, which is,

$$R = R_{dc}\left(1 + \frac{\omega^2 l^2 \times 10^{-19}}{1.2\,R^2}\right) \tag{9-2-1}$$

where R/l is the resistance of the wire of the resistor (if wire-wound) in ohms per centimeter. This type of loss is noticeable only at low resistance values if rather thick, low-resistance wire is used.

If ferromagnetic materials are brought near the resistor, hysteresis loss will decrease the effective value of R_e. An extreme case is the ac resistance of an iron-core coil (see Sec. 9-2-3).

Resistor Phase Angle. The resistor phase angle, or its tangent Q, can be calculated by using the same equivalent circuit. Generally the lumped capacitance and inductance are the dominant factors, giving a value of Q of

$$Q = \tan\theta = \omega\left(\frac{L}{R} - RC\right) - \frac{\omega^3 L^2 C}{R} \tag{9-2-2}$$

The last resonant term can usually be neglected.

Standard Resistors. A good standard resistor is one that displays minimum change on account of the causes mentioned above and, even more important, is very stable with time. Standard resistors up to 10 MΩ are usually wire-wound. Stability is improved by low-tension winding, heat cycling, and sealing in a chemically inactive oil or gas.

The Thomas 1-Ω resistor, which has been the best available standard for a long time, is bifilar wound of heavy manganin wire in a sealed container. It has four terminals brought out so that any four-terminal measurement will be independent of the resistance of the connecting leads, the terminals, and the contact between them (see Secs. 9-1 and 9-4).

New 10-kΩ standards are coming into use. These use Evanohm wire which can be treated to have a temperature coefficient of less than 0.2 ppm/°C over a narrow temperature range. While these standards are also four-terminal, lead resistance is much less critical at this higher resistance value. However, at this resistance level, shunt leakage resistance must be kept very high and a guarded measurement is recommended.

The range of resistance standards extends down to 10 μΩ at which level precision four-terminal shunts used for high-current measurement have an accuracy of 0.04 percent. While wire-wound resistors over 100 MΩ have been made, film types are usually used in this range and on up to 10^{13} Ω. Sometimes T networks are used as high-valued three-terminal resistance standards because their transfer resistances can be very high (see YΔ transformation, Sec. 9-1).

9-2-2 *Capacitors* [6, 7]

Behavior of Capacitors. There are many types of capacitors with such widely different characteristics that a tabulation of some of their important specifications is useful in interpreting measured values and selecting types for actual use (see Table 9-2). The specifications given are intended to be typical. Actual specifications can have a wide spread and can depend greatly on the capacitance value.

Frequency Characteristics of Capacitors. A rather complex equivalent circuit is shown in Fig. 9-6, which represents a variety of phenomena which affect the value of capacitance and D over a wide frequency range.

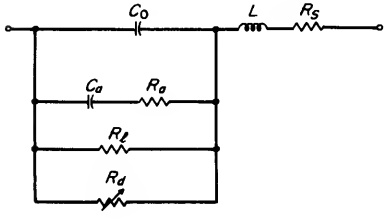


FIG 9-6 An equivalent circuit of a capacitor.

TABLE 9-2 Typical Specifications of Various Types of Capacitors

Type	Usual capacitance range, μF	Best tolerance, %	TC, ppm/°C	D, 1 kHz	Approx. voltage recovery,† %	Insulation resistance, 25°C, Ω-F
Polystyrene	0.001–1	½	−120	0.0005	0.1	$>10^6$
Polycarbonate	0.001–10	1	±100	0.002	0.2	$>10^5$
Mylar	0.001–10	1	Nonlinear	0.01	0.5	$>10^5$
Paper, impregnated	0.001–100	1	±500	0.005	5	$>10^4$
Mica	1 pF–0.02	¼	±100	0.001	3	$>10^4$
Ceramic, low k	0.0001–0.01	5	0 to −750	0.001	2	>1,000
Ceramic, high k	0.001–0.5	20	10,000‡	0.01	4	>700
Tantalum electrolytics	0.005–100	10	1,000	0.04	2	>100
Aluminum electrolytics	5–100,000	20	10,000‡	0.05‡	10	>30‡

† 10-second discharge.
‡ Varies widely.

This circuit is more useful when simplified to apply to a specific frequency range and type of capacitor.

The electrostatic, or dc, value of capacitance is C_0 plus the relatively small capacitance C_a, which represents the increase in effective dc capacitance over its audio-frequency value resulting from *interfacial polarization* [8]. This effect can be considered a gradual redistribution of charge throughout the dielectric and is most pronounced in composite dielectrics. It causes what is called *dielectric absorption*, characterized by the additional charging current flowing long after an ideal capacitor would have been charged and the gradual recovery of voltage after the capacitor has been momentarily short circuited. The R_aC_a time constant may be many seconds, hours, or even days. Table 9-2 gives typical values of *voltage recovery* after a 10-sec discharge which is approximately equal to C_a/C_0 if the time constant is very long.

Capacitors with dielectrics having dipole moments have a *molecular polarization* in the rf range which causes a further change in capacitance. This could be represented by a second RC combination having an appropriate time constant.

At high frequencies the series inductance increases the effective value of capacitance sharply as series resonance is approached, as shown by Eq. (9-2-3)

$$C_{rf} = \frac{C_0}{1 - \omega^2 LC_0} = \frac{C_0}{1 - (f/f_r)^2} \tag{9-2-3}$$

where f_r is the resonant frequency $1/(2\pi\sqrt{LC})$.

There are several causes of energy loss in most types of capacitors. At dc all capacitors have a finite *leakage resistance* R_l (also called *insulation*

resistance), even though in some cases it may have an extremely high value. Ideally, the product R_lC is independent of value of capacitance for a given type of capacitor, and therefore the product is specified (see Table 9-2). However, surface leakage decreases this quantity for low-valued capacitors. Measuring the true value of R_l can be exasperating if appreciable dielectric absorption is present, and the apparent leakage resistance after a given time (1 or 2 min) is used as a specification.

Leakage resistance is only a minor cause of dielectric loss, except at extremely low frequencies. Likewise, except at high frequencies (or for very large capacitors), series resistance of the leads, plates, or foils makes but a small contribution to the total value of D. The major part of D is energy loss in the dielectric material itself. Ideally, this could be explained by the polarization effects which would make well-defined humps in a D versus frequency plot. While such humps are often present, they tend to be spread out as a result of nonhomogeneity in the dielectric structure and in the applied electric field. There is also a residual loss present at all frequencies. The variable resistor R_d is used to represent these miscellaneous effects.

In simpler models, all dielectric loss could be represented by a single variable resistor R_d. Except over a narrow frequency range, however, the use of a single fixed series or parallel resistance to simulate all loss can be very misleading (see Sec. 9-1). The often used equivalent series resistance is much larger than the actual physical series resistance and therefore is a particularly poor representation at frequencies much higher than the one at which it was determined.

Capacitance Standards. The type of capacitor used as a standard depends greatly upon the capacitance value. The ultimate standard, the Thompson-Lampard calculable capacitor used for the absolute determination of the ohm (see Sec. 9-7), is practical only for values near 1 pF. Such capacitors are very expensive and are kept by only a few national laboratories. The most stable working standards are three-terminal 10-pF capacitors, designed by the Bureau of Standards, that use fused silica disks as the dielectric. Even though their temperature coefficient is relatively high (about 11 ppm/°C), their stability and ruggedness make them excellent for use as traveling standards for interlaboratory comparisons. These are now available commercially.

Other commercially available standards are of the parallel-plate type and are made entirely of Invar steel for a low-temperature coefficient. They are sealed and filled with an inert gas to make their value independent of humidity and atmosphere pressure. These capacitors are also three terminal, have values up to 1,000 pF, and are stable to a very few ppm per year. Larger capacitance standards up to 1 μF use high-quality mica as a dielectric, with deposited silver electrodes. They are specified

at better than 100 ppm/year stability. Standards of lower stability and accuracy, using a variety of constructions, extend the range down to 0.001 pF and up to 1 F.

9-2-3 *Inductors* [9, 10]

The value of an inductor depends more on the conditions of measurement than does that of the other types of component. This is particularly true for iron-core inductors, for which measurement repeatability within a few percent is considered good and is usually adequate.

Because many inductors have a rather low Q, the difference between the effective series and parallel values is apt to be considerable. Unless specified otherwise, the series value is generally used. While this is reasonable for air-core inductors, the parallel value is a more meaningful specification for iron core inductors.

Iron-core inductors are nonlinear, and therefore the measured inductance depends on the level of the test signal and the level of any dc bias. The combination of ac and dc used should simulate that of the actual application if possible. Two widely used plots are inductance versus ac level with no dc applied and the incremental inductance plot of inductance versus dc bias current with very small ac signal (see Sec. 9-5-4). If series inductance is plotted, the resulting curves can be difficult to explain because the core losses, which behave as parallel resistances, vary widely and change the effective series inductance value (see Table 9-1).

The initial condition of the core will affect the measured value, for it may have been left in a state of residual magnetization. It should be demagnetized by applying a large saturating ac signal and slowly reducing this to zero.

Air-core inductors of the solenoid type are affected by the proximity of conducting material. This is most noticeable in rf coils where the Q can be greatly reduced by shield cans.

Frequency Characteristics of Inductors. The equivalent circuit of Fig. 9-7 is usually adequate for both air- and iron-core inductors, but R_h, which represents hysteresis loss, is removed for air-core types, and R_e, which represents eddy-current losses, is much higher.

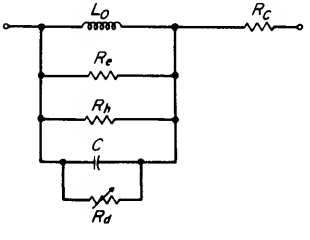


FIG 9-7 An equivalent circuit of an inductor.

The series resistance R_c, which represents the actual winding resistance, does not affect the effective series value of inductance. This resistance is equal to the actual dc winding resistance at low frequencies, but increases in the rf range because of skin and proximity effects. For air-core inductors R_e represents these effects and the loss in a shield or other nearby conductor as well.

In iron-core devices, R_e also represents the eddy-current loss in the core. This resistance value depends on the resistivity and dimensions of the core laminations or magnetic particles, but is nearly independent of level and frequency. The parallel resistance R_h represents hysteresis loss. It has a value proportional to frequency and to the flux density raised to some fractional power, depending on the Steinmetz exponent [11].

All coils have capacitance that increases the effective inductance through resonance. This capacitance is actually distributed and can result in several resonances at different frequencies which produces a wide variation in effective inductance. The loss in this distributed capacitance R_d can be an important cause of reduction in the Q of air-core coils at high frequency.

Standard Inductors. Precision standard inductors are usually wound as toroids or solenoids on ceramic cores. While solenoids generally have a higher Q, toroids are less affected by nearby metal. Two solenoids connected in series and placed side by side can have sufficient mutual coupling to change the total value appreciably.

The position of connecting leads can greatly affect the measured value of low-valued inductors [12]. Such standards are sometimes made with a shorting link at their terminals. The calibrated value is the difference measured when the short is removed.

Lower-accuracy standards and decades use powdered iron or ferrite cores to improve the Q. They are level sensitive, the calibrated value usually being the value extrapolated to zero signal level and zero frequency.

9-3 Meter Methods to Measure Impedance

9-3-1 Direct-current Meters

Ammeter-Voltmeter Method. The most obvious way to measure resistance is to measure voltage and current separately and calculate resistance from their ratio. This ammeter-voltmeter method, Fig. 9-8, is rarely used in instruments because it depends on the accuracy of two meters and it requires a calculation. Also, both connections shown are subject to error, for in the first the voltage measured includes the voltage drop in the ammeter, and in the second the current measured includes the

current in the voltmeter. It is not difficult to make corrections for these errors, and electronic meters, which draw extremely low power from the measurement circuit, have negligible error from these sources, except when measuring extreme values of resistance.

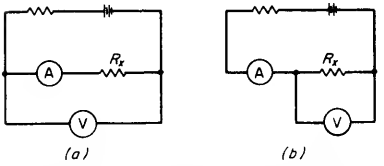


FIG 9-8 The ammeter-voltmeter method for measuring resistance.

Ohmmeters. Rather than measure both voltage and current, many instruments keep one constant so that the measurement of the other quantity is a measurement of resistance. If the current through a resistance is kept constant, a voltmeter across it reads resistance linearly. This is the principle of the digital ohmmeter, which is made by simply adding a regulated current source to a (digital voltmeter) DVM.

If the voltage is kept constant, an ammeter in series would have deflection proportional to conductance, but the meter is often calibrated in terms of resistance, going to ∞ at 0 current. The resulting scale is nonlinear but practical as long as a digital indication is not required.

Actually, ohmmeters are usually designed to have nonlinear scales which extend from 0 to ∞ so that their range is increased and the range switch does not have to be used as often. The most common ohmmeter circuit is shown in Fig. 9-9. In this circuit R_1 is adjusted to give a zero reading when R_x is shorted, and R_2 is a meter shunt used to change the range.

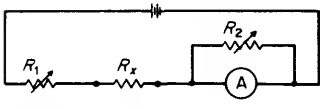


FIG 9-9 A practical ohmmeter circuit.

While digital ohmmeters can be quite accurate and 0.01 percent is not uncommon, simple ohmmeters incorporated in multimeter testers are limited in accuracy to several percent.

The ohmmeter principle can be extended to very low resistance values

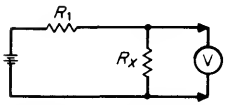


FIG 9-10 A milliohmmeter circuit $R_1 \gg R_x$.

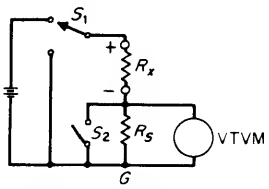


FIG 9-11 The basic megohmmeter circuit.

if a sensitive voltmeter is available (Fig. 9-10). These meters can make a four-terminal resistance measurement (see Sec. 9-1) if the resistance of the voltmeter is much larger than R_x or the resistance of the leads.

Megohmmeters. A very common circuit for measuring high-valued resistors is the megohmmeter of Fig. 9-11. Here the "ammeter" consists of a high-resistance standard shunted by a vacuum-tube (or FET) voltmeter whose deflection is proportional to $R_s/(R_s + R_x)$. Such instruments measure resistances up to 10^{12} Ω or more, but at relatively low accuracy.

Comparison Methods. If an unknown resistor is placed in series with a standard resistor of approximately equal value, the ratio of the voltages across them is also the ratio of their resistances. Two measurements are necessary. The closer in value the two resistors are, the more accurate the measurement is because meter nonlinearity and meter loading will cause almost equal errors in both measurements and therefore cause only small error in their ratio.

A related method, the potentiometric method (Fig. 9-12), uses a null detector and a potentiometer to measure the voltage ratio. The potentiometer measures R_x and R_s in turn. This method can be extremely accurate if the voltage supplies are stable and a high-resolution potentiometer is used. Also, because no current flows at balance, there is no error caused by lead resistance.

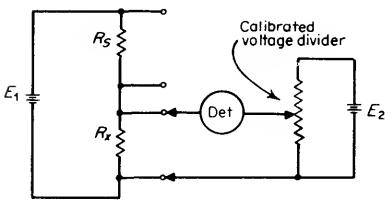


FIG 9-12 The potentiometric method of comparing two resistors.

9-3-2 *Capacitance and Inductance Meters*

Ammeter-Voltmeter-Wattmeter Method. If the ammeter-voltmeter method is used for ac, the ratio of the meter readings is the *magnitude* of impedance and admittance. To separate resistance and reactance, some type of phase-measuring or discriminating circuit is necessary. The

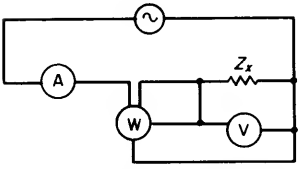


FIG 9-13 The ammeter-voltmeter-wattmeter method for measuring impedance.

classical method is to add a wattmeter to the ammeter-voltmeter circuit as shown in Fig. 9-13. The wattmeter reading is proportional to $|I|\ |E|\ \cos\theta$, where θ is the phase angle of the impedance. Because the other two meters measure $|I|$ and $|E|$, the angle θ may be determined and both R and X calculated.

Alternating-current Comparison Methods. The dc method of comparing the voltages across a standard and an unknown resistor in series can be easily extended to ac if only magnitude information is desired. If phase information is also required, a phasemeter can be used or an oscilloscope with differential vertical and horizontal inputs can give an elliptical trace from which phase angle can be calculated [13].

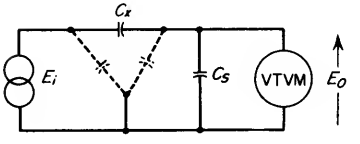


FIG 9-14 A capacitance meter circuit.

Capacitance Meter. The simple divider circuit of Fig. 9-14 is convenient for measuring nearly perfect capacitors. Here,

$$\frac{E_o}{E_i} = \frac{C_x}{C_x + C_s}$$

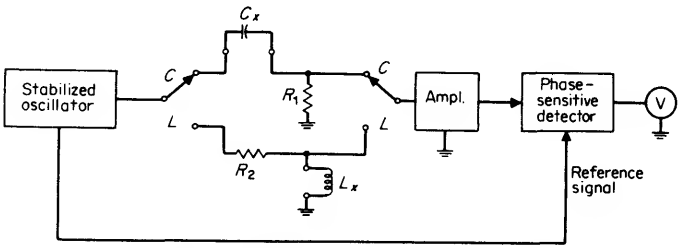


FIG 9-15 A capacitance-inductance meter (Boonton Electronics Corporation 71A).

so that the meter reading may be calibrated to read C_x, and C_s may be switched to change range. Three-terminal measurements are possible if C_s is large, because stray capacitance shunting C_s would then have little effect.

Capacitance-Inductance Meter (Boonton Electronics Corporation 71A or 71D). The capacitance meter circuit of this instrument, Fig. 9-15, uses a resistor to sample current, rather than the capacitor of the previous instrument, because it operates at 1 MHz, where inductance in series with a large capacitor would cause appreciable error. It also uses a phase-sensitive detector which indicates the component of measured voltage in quadrature with the input. This greatly reduces errors caused by loss in the capacitor or reactance in the resistor.

An RL circuit is used to measure inductance, with the positions of the standard resistor and the unknown interchanged so that the deflection is proportional to inductance and not its reciprocal.

Frequency-deviation Capacitance Meters. Several instruments have been designed in which an unknown capacitance changes the frequency of an oscillator and this frequency deviation is used to measure capacitance. Resonant circuits can be used to detect capacitance limits for sorting applications. (Refer to Sec. 9-6-4 for discussion of the Q meter and a resonant capacitance meter.)

9-3-3 Complex Impedance Meters

If a known and constant current is passed through an unknown impedance, the in-phase and quadrature components of the resulting voltage are measures of the resistance and reactance of the unknown. Alternatively the magnitude and phase angle of the voltage may be measured to obtain impedance magnitude and phase. Likewise a known and constant voltage may be applied and the resulting current measured to get the corresponding admittance quantities. This simple principle has been the basis for several impedance meters that characterize the unknown

impedance completely, instead of giving just one component as do the previously described instruments. The accuracy and range of these instruments are dependent upon the electronic circuits used to control the input signal and measure the output quantities, and upon the output-indicating device. Recently, precision high-feedback circuits and digital voltage indicators have made great accuracy possible in such circuits, although they still are less accurate than dc digital ohmmeters.

Vector Impedance Meter (Hewlett-Packard 4800A) [14]. This instrument measures magnitude and phase in the manner described above over wide ranges of impedance (1 Ω to 10 MΩ) and frequency (5 Hz to 500 kHz). For impedance up to 1 kΩ, the input current is constant and the output voltage magnitude is used to indicate impedance magnitude. Above 1 kΩ, voltage is applied and current is measured, so that the output signal actually is proportional to admittance magnitude. However, the indicating meter has an inverse scale which reads impedance magnitude on these higher ranges. The impedance phase angle is the phase angle between voltage and current no matter which is input and which is output.

An impedance meter operating at a frequency that is a power of $10 \times 1/2\pi$ will read $L(1 + 1/Q^2)^{1/2}$ and thus be direct reading in inductance if Q is high. Because this instrument actually reads both $|Z|$ and $|Y|$, it can also read $C(1 + D^2)^{1/2}$ at these frequencies.

9-3-4 *Resistance and Impedance Comparators*

Direct-current Comparators (Limit Bridges). Small differences in resistance from some nominal value may be measured by applying equal voltages of opposite polarity to series-connected standard and unknown resistors, as shown in Fig. 9-16*a*. One way to provide virtually equal voltages is

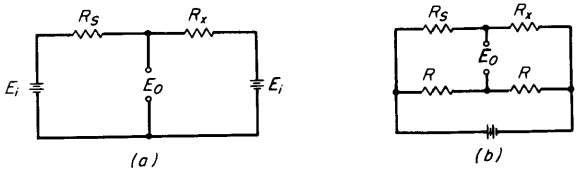


FIG 9-16 Direct-current resistance comparators.

to use a precision resistance divider to form the bridge circuit as in Fig. 9-16*b*. The output voltage as measured by a high-impedance voltmeter is proportional to

$$\frac{R_x - R_S}{(R_x + R_S)/2} = \frac{\Delta R}{R_{av}} \tag{9-3-1}$$

If R_s is adjusted to the nominal value, the voltage reading is the difference in the unknown from the nominal value as a percent of the average of the standard and unknown. A more desirable indication would be the deviation as a percent of the standard. For small deviations, the two functions are almost equal, and for larger deviations a nonlinear meter scale is often used. An alternative method of getting $\Delta R/R_S$ is to add the output voltage to the voltage supplied to R_s.

This circuit is often used to make rapid production or inspection tests on large batches of resistors of the same nominal value. Voltage comparators are often used to detect tolerance limits and to provide a go–no-go indication or to actuate a rejecting mechanism in automatic systems.

In using comparators, one must distinguish between the accuracy of the meter reading and the accuracy of the comparison. Even if the meter has low accuracy, precise comparisons can still be made if the resistance difference is small. For example, a 3 percent measurement of a resistance difference of 0.1 percent is a comparison of 0.003 percent.

Alternating-current Comparators [15]. This same principle can be used for impedance, but the resulting unbalance voltage is now complex. If phase-sensitive detectors are used to separate this output voltage into components that are in phase (real) and in quadrature (imaginary) with the input signal, the two outputs are proportional to

$$\operatorname{Re}\frac{E_o}{E_i} = \frac{(|Z_x| - |Z_S|)/(|Z_x| + |Z_S|)}{1 + [\cos(\theta_x - \theta_S) - 1]/(1 + |Z_x|/2|Z_S| + |Z_S|/2|Z_x|)}$$

and

$$\operatorname{Im}\frac{E_o}{E_i} = \frac{\sin(\theta_x - \theta_s)}{\cos(\theta_x - \theta_S) + |Z_x|/2|Z_S| + |Z_S|/2|Z_x|} \tag{9-3-2}$$

If the impedance-magnitude deviation is less than 10 percent and the phase difference less than 0.1 rad, these reduce to $(|Z_x| - |Z_S|)/(|Z_x| + |Z_S|)$ and $\theta_x - \theta_S$ in radians with an error of less than 0.25 percent *of reading*. Moreover, if Z_x and Z_S are both relatively pure reactances, the magnitude unbalance closely approximates the C, R, or L unbalance and the θ unbalance becomes a D or Q unbalance (whichever is the smaller quantity). Alternating-current comparators usually have inductively coupled ratio arms (see Sec. 9-5-3).

9-4 Direct-current Bridges

9-4-1 The Wheatstone Bridge

The relation between a meter and a bridge method of measuring impedance is analogous to that between a spring scale and an analytical

balance for measuring mass. The spring scale and meter methods depend on the calibration, linearity, and stability of a measuring device. The balance and bridge depend on the calibration and stability of a passive quantity similar to that being measured and on the sensitivity of detecting the difference between them. Just as a mass standard can be more stable than the calibration of a spring, so an impedance standard can be more stable than the gain of an amplifier or an indicating meter movement.

History. Hague [16] credits the bridge principle to S. H. Christie, referring to a paper in 1833, but credits Sir Charles Wheatstone with drawing attention to Christie's idea and with the first application to the comparison of resistance. Wheatstone called his circuit (Fig. 9-17) a *resistance balance,* and the branches he called the *arms,* both terms drawn from the analogy used above. Wheatstone compared only equal resistors and, therefore, had equal ratio arms. Werner von Siemans was the first to use unequal arms and thus made balances possible between resistances of widely differing values.

The term *bridge* was first applied to the detector branch of the circuit because points A and B were *bridged* by the galvanometer. Later the term *bridge* was used to refer to the whole circuit.

Modern Wheatstone bridges use this same circuit, the main improvements being in the range and quality of the resistors and the design of the source and detector. Most modern bridges have one fixed arm, usually the one opposite the unknown resistor; one variable arm, which is calibrated to give the value of the unknown; and one arm that is switched in decade values to provide a multiplier on the indicated value.

The Balance Equation. There is no current flowing in the galvanometer in the basic bridge circuit of Fig. 9-17 if

$$\frac{R_3}{R_1} = \frac{R_4}{R_2} \quad \text{or} \quad R_4 = \frac{R_2 R_3}{R_1} \tag{9-4-1}$$

Several important properties of the bridge circuit may be deduced from this simple equation:

1. The balance, or null, condition is independent of the magnitude of

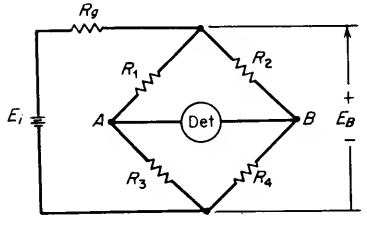


FIG 9-17 The Wheatstone bridge.

the input voltage or its source impedance. These quantities do not appear in the expression.

2. The balance condition is independent of the sensitivity of the detector, the impedance of the detector, or any impedance shunting the detector. These factors influence the ability to find a null, but not the null condition itself.

3. The balance condition is unchanged if the source and detector are interchanged. Interchanging source and detector is the same as interchanging R_2 and R_3, and it is the product of the two resistors which must remain unchanged.

4. A given resistor (R_1, for example) is proportional to the resistance of the adjacent arms (R_2 and R_3) and inversely proportional to the resistance (or proportional to the conductance) of the opposite arm (R_4).

Sensitivity. The sensitivity of a bridge can be expressed in many ways by using output-to-input ratios of voltage, current, or power or any combination. The most useful expression for any given application is the one that uses the quantities that represent the actual limitations of applied power and detector sensitivity. A general expression for voltage ratio near balance is

$$\frac{E_o}{E_i} = \frac{\delta R_d}{R_1 + R_2 + R_3 + R_4 + R_d\,(2 + R_1/R_3 + R_3/R_1) + R_g(2 + R_1/R_2 + R_2/R_1) + R_d R_g(1/R_1 + 1/R_2 + 1/R_3 + 1/R_4)} \tag{9-4-2}$$

where δ is the fractional bridge unbalance, that is, $R_4 = (1 + \delta)R_2R_3/R_1$; R_d is the detector resistance; R_g is the source resistance; and E_i is the open-circuit source voltage (see Fig. 9-17).

The ratio of E_o/E_B, where E_B is the actual voltage across the bridge, can be easily obtained from Eq. (9-4-2) by setting R_g equal to zero. The open-circuit output voltage and short-circuit output current can be found by setting R_d equal to infinity and zero respectively.

A common situation is a very low-impedance source and a very high-impedance detector, in which case the expression is very simple and gives maximum output voltage for a 1:1 ratio bridge $R_1 = R_3$. Maximum output power for a given power in the bridge is achieved when the resistances of the detector and all bridge arms are equal. If the bridge arms are unequal, reversing the source and detector connections will give increased power out if it results in a better impedance match between the detector and the bridge output resistance, which, near balance, is

$$\frac{R_1(1 + R_2/R_1)}{1 + R_1/R_3}$$

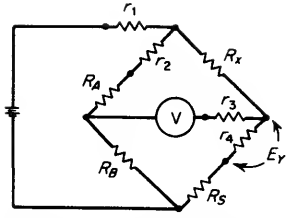


FIG 9-18 A four-terminal resistor measured on a Wheatstone bridge.

9-4-2 *Measurement of Low-valued Resistors*

The measurement of low-valued, two-terminal resistors is affected by the resistance of the connecting leads and terminals, so that low-valued standards are made with four terminals to allow repeatable measurements. Such four-terminal resistors can be measured with a Wheatstone bridge, but will be subject to some error. In Fig. 9-18, a four-terminal resistor is connected to a Wheatstone bridge with four leads having resistances r_1, r_2, r_3, and r_4. It is easy to see that r_1 and r_3 have no effect on the balance condition, but r_2 and r_4 would make the bridge read low by approximately $(r_2/R_A + r_4/R_S)100$ percent. If R_A and R_S are both larger than R_x, this error is preferable to the error that would result if the supply were connected to r_2 and the detector to r_4, for these resistances would appear in the R_x arm.

Although it is possible to have both R_A and R_S larger than R_x, it is not usually practical to have then both much larger because this requires a very large value of R_B to satisfy the balance equations and gives poor sensitivity. A common measurement in precision work is the comparison of resistors of equal value. This circuit has no advantage for such measurements.

However, if $R_S = R_x$, one solution is to take the average of two measurements, one with the detector tied to r_3 and the other with it tied to r_4. This puts the undesired voltage drop E_Y first in one arm and then in the other. This gives a resulting error of approximately $(r_4/R_x - r_4/R_S)100$ percent.

The Kelvin Double Bridge [17]. An alternative method, devised by William Thomson, Lord Kelvin, divides this undesired voltage drop E_Y proportionately between the two arms to avoid error. Figure 9-19 shows his bridge in comparing two four-terminal resistors R_x and R_S. Here lead resistances r_1 and r_8 are in series with the source and cause no error. Resistances r_2 and r_7 together cause an error of approximately

$$(r_7/R_B - r_2/R_A)100 \text{ percent}$$

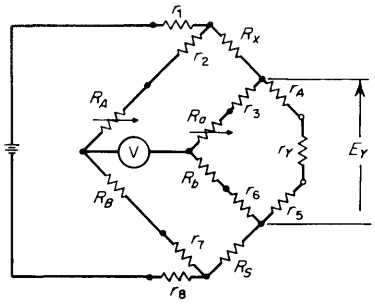


FIG 9-19 The Kelvin double bridge.

which is small if R_A and R_B are large or if this expression is made equal to zero by adjustment of r_2 or r_7 (called a *lead compensation* adjustment).

The voltage E_Y is caused by the IR drop in r_4 and r_5 and in resistance r_Y connected between them (called the *yoke*). If the ratio of the additional bridge arms R_a/R_b is equal to R_x/R_S (and r_3 and r_6 are zero), then there is no net error because the voltages across R_a and R_b cause equal and opposite errors which cancel. If r_3 and r_6 are not negligible, the error expression is very nearly

$$\frac{r_4 + r_5 + r_Y}{R_S} \frac{(R_b + r_6)R_A - (R_a + r_3)R_B}{R_A(R_a + R_b)} 100 \text{ percent} \qquad (9\text{-}4\text{-}3)$$

The second factor is kept small by ganging the adjustments of R_a and R_A so that, very nearly, $R_a/R_b = R_A/R_B$. Errors that are due to inaccuracies in tracking and to r_6 and r_3 are further reduced by a separate *yoke adjustment* (usually in R_b) which is made when the yoke connection is broken ($r_Y = \infty$).

The lead compensation adjustment is much more important than the yoke adjustment as long as $R_S \gg r_4 + r_5 + r_Y$. If low-valued ratio arms are used, the lead resistances r_2 and r_7 are critical. This suggests the use of additional arms to balance the resulting voltage drops. A bridge with six pairs of arms [18] is shown in Fig. 9-20 with the lead resistance omitted.

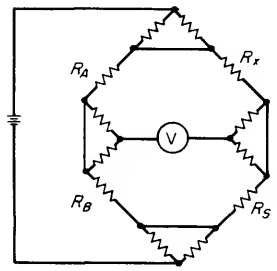


FIG 9-20 Warshawsky's resistance bridge.

9-4-3 *Measurement of High-valued Resistance*

Megohm Bridges. The Wheatstone bridge is used for the measurement of resistors up into the 10^{15}-Ω range and higher by appropriate choice of values in the bridge arms and by using high-impedance "electrometer" types of dc amplifiers as detectors. In choosing the values for the arms of such a bridge, one has the choice of either using very high valued resistors, which are generally not very stable, or using lower-valued, more stable resistors and an extreme bridge ratio, which results in a great loss in sensitivity. The variable element usually cannot have an excessively high value, particularly if it is a rheostat. Therefore, this element is placed in the opposite arm (from the unknown) and the unknown resistance is proportional to its conductance. This also has the advantage of extending the resistance range to ∞ as the resistance of this adjustment goes to zero.

If very high valued resistors are used for the range selection, they are often made adjustable and a calibration procedure is used before making precision measurements to ensure accuracy. This can be either a comparison of measurements on each range against the next lower range by using an external resistor which can be measured on both ranges, or a comparison against a lower-valued standard by connecting each standard resistor in turn to the unknown arm.

The bridge ratio can be extended by use of additional bridge arms R_C and R_D, as shown in Fig. 9-21, which permits the use of resistors of more practical value. This modification can be considered two ways: as a divider across R_B placing a very small voltage across R_S, or as a T network R_D, R_C, and R_S placed in the R_S arm whose equivalent value, from the YΔ transformation, is very large. No matter how it is looked at, it results in a loss in bridge sensitivity.

Guarding. Just as low-resistance measurements are affected by series lead impedance, high-resistance measurements are affected by shunt-leakage resistance. The two problems and their solutions are analogous (duals). The leakage resistance appears between the leads themselves, between the terminals of the bridge and its internal parts, and across the unknown resistor itself. High-valued resistance standards are made

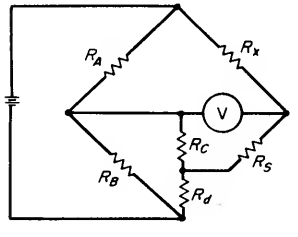


FIG 9-21 A megohm bridge with additional arms to increase the ratio.

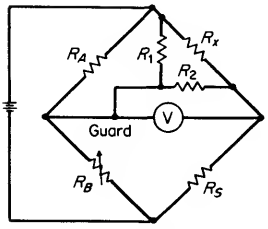


FIG 9-22 A three-terminal resistor measured on a Wheatstone bridge.

with three-terminals, and a three-terminal measurement can be made to avoid this type of error.

The simplest connection for this third (guard) terminal is to a bridge corner such that the shunt resistances are placed across bridge arms with low resistance. In Fig. 9-22, R_2 causes no error (only sensitivity loss) and R_1 causes an error of approximately $(R_A/R_1)100$ percent.

If this guard connection were not made, the error would be approximately $R_x/(R_1 + R_2)100$ percent. The error is greatly reduced only if $R_A \ll R_x$.

The Wagner Guard Circuit. By adding two additional arms to the bridge, the error caused by the shunt resistance to the guard can be removed (Fig. 9-23). In this circuit, if the ratio of R_4 to the parallel combination of R_1 and R_3 is equal to R_S/R_x, there is no voltage across R_2, so that no current will flow through it and it can cause no error. This condition is met by making a preliminary or *Wagner guard* balance with the detector connected between points B and C.

The extra guard arms can also be added across the bridge from A to C, where their ratio should be that of R_A to R_x. Putting arms in both directions, tied in the middle, results in a guard whose balance is much less critical, but requires two preliminary adjustments [19].

9-5 *Low-frequency Bridges*

9-5-1 *General*

History [16]. In 1865, Maxwell used a ballistic galvanometer to measure the transient caused by breaking the battery connection to a Wheatstone bridge when one arm contained an inductor. Later he balanced this transient by adding a second, variable coil to the bridge. This bridge

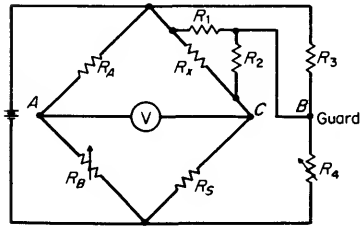


FIG 9-23 The Wagner guard circuit on a Wheatstone bridge.

required two adjustments: one for the steady-state condition, the resistance balance, and one for the transient, the ballistic or inductive balance. Ayrton and Perry used ganged commutators to reverse both the battery and detector connections. This not only provided a continuous series of transients, it also kept the galvanometer deflection in the same direction for a given bridge unbalance.

The advent of the telephone receiver in 1875 made available a much more sensitive detector. Several famous experimenters used it with an induction-coil source to develop bridge circuits. Wien used a vibrator and later an alternator to get a test signal of constant frequency. Developing work by Overbeck, he described the principles of ac bridges in 1891. The most important work since then is probably that of Blumlein in the 1920s describing the advantages of inductively coupled ratio arms, which did not come into wide use until much later.

Ratio and Product Bridges. If we substitute impedances for the resistances in the Wheatstone-bridge balance equation, we have

$$Z_x = Z_1 \frac{Z_2}{Z_3} = Z_1 Z_2 \frac{1}{Z_3} = Z_1 Z_2 Y_3 \tag{9-5-1}$$

This equation is written in three ways to emphasize the properties of ratio and inversion. If the ratio Z_2/Z_3 is real, the equation shows the *comparison* of the unknown impedance with one that is similar but not necessarily of the same magnitude. If the *product* Z_1Z_2 is real, the equation shows the proportionality between an impedance and an admittance so that we could have (arm x is opposite arm 3)

$$R_x = R_1 R_2 G_3 \qquad \text{or} \qquad L_x = R_1 R_2 C_3 \tag{9-5-2}$$

This *inverting* ability is particularly useful for inductance bridges because inductors are poor bridge components (see Sec. 9-2).

Series and Parallel Bridges. In general, the unknown impedance will be complex, and therefore, for a null balance, the right-hand side of the equation must also be complex with real and imaginary parts equal to those of the unknown. While all three impedances of the right-hand side could be complex (and, of course, are to some degree, except in theory), usually only one is made complex in order to simplify the balance condition. If either of the adjacent arms Z_1 or Z_2 is a series combination of a resistance and a reactance, we have

$$R_x + jX_x = (R_1 + jX_1) \frac{Z_2}{Z_3} = R_1 \frac{Z_2}{Z_3} + jX_1 \frac{Z_2}{Z_3} \tag{9-5-3}$$

Such a bridge would indicate the series impedance components if the readout were proportional to these two terms. A bridge also reads series

impedance if the opposite arm Z_3 is a parallel combination, for in this case $1/Z_3 = Y_3 = G_3 + jB_3$, so that $R_x + jX_x = (G_3 + jB_3)Z_1Z_2$. Likewise, it is easy to show that a parallel combination in an arm adjacent to the unknown or a series combination in the opposite arm will result in a bridge that indicates parallel components, for the balance equation can also be written as $Y_x = G_x + jB_x = Y_1Y_2/Y_3 = Y_1Y_2Z_3$.

Position of the Variable Components. In order to make the real and imaginary parts of the balance equation equal, two components of the bridge must be adjustable. The position of these variable components in the circuit determines what quantities the bridge will measure because each bridge readout is (almost always) proportional, or inversely proportional, to the value of one adjustable component. There are three possibilities:

1. *Both adjustments in the same arm.* In this case the bridge reads the real and imaginary parts separately, such as R and X or L and R.

Example

$$R_x + j\omega L_x = R_1R_2(\dot{G}_3 + j\omega\dot{C}_3) = R_1R_2\dot{G}_3 + j\omega R_1R_2\dot{G}_3$$

where the dots indicate the variable components with calibrated dials, or other readouts.

2. *Adjustments in different arms.* This type of bridge reads the value of one complex component and one of D, Q, or phase angle.

Example

$$\begin{aligned} R_x + j\omega L_1 &= \dot{R}_1R_2(\dot{G}_3 + j\omega C_3) \\ L_x &= \dot{R}_1R_2C_3 \\ Q &= \frac{\omega L_x}{R_x} = \frac{\omega C_3}{\dot{G}_3} = \omega C_3\dot{R}_3 \end{aligned}$$

To read angle instead of Q, the indicating scale of R_3 would be a nonlinear function of its resistance. If C_3 were adjustable rather than G_3, the above bridge would read R and Q.

3. *One adjustment (or both) in the unknown arm.* This is a *substitution* measurement in which the *change* in the adjustment when the unknown is added is equal to the similar part of the unknown.

Example

$$\begin{aligned} R_x + j\omega L_x + \dot{R}_Y &= R_1R_2(G_3 + j\omega\dot{C}_3) \\ R_x &= R_1R_2G_3 - \dot{R}_Y \\ L_x &= R_1R_2\dot{C}_3 \end{aligned}$$

Bridge Residuals. All arms of all actual bridges contain some resistance, inductance, and capacitance, and as a result their actual balance condi-

tions are very complicated functions. The errors resulting from these residual parameters limit the accuracy and frequency range of the bridge. Extensive shielding and compensating techniques are used in many commercial bridges.

9-5-2 *Classification of Four-arm Bridges* [20, 21]

While many bridges are possible, the list can be greatly reduced if one:

1. Does not indicate which components are adjustable.
2. Does not use inductors as bridge elements (except in the basic inductance comparison bridge).
3. Assumes all ratio arms in ratio (comparison) bridges are resistors.
4. Does not indicate the position of source and detector.
5. Does not list all obvious parallel versions of bridges whose series versions are shown (some common ones are shown).

The bridges remaining, their balance equations, and comments are listed in Table 9-3.

Frequency Bridges. The R, C, and L balances of the bridges of Table 9-3 (except for the resonant bridge) are all independent of frequency simply because in the calculations it was assumed that the unknown was either a series or parallel combination, depending on which would be indicated directly by the adjustments (Sec. 9-5-1). There is a large family of bridges that are frequency dependent and can be used to determine frequency. In these *frequency bridges* all four arms are specified and, therefore, there is no "unknown" impedance. The best known of these bridges is the Wien bridge used in many RC oscillators. This can be considered a series RC comparison bridge measuring a parallel RC combination, or vice versa.

9-5-3 *Bridges with Inductively Coupled Ratio Arms* [22, 23]

Equivalent Circuit of a Tapped Coil. Ratio, or comparison, bridges can use any similar impedances as ratio arms, but inductors would be the worst choice because of their low Q, poor frequency characteristic, nonlinearity, vulnerability to pickup, and size. However, if two inductors are tightly coupled, they become the best choice, for they have two unique advantages, excellent ratio and immunity to loading, which are explained by the low-frequency equivalent circuit for such a transformer as shown in Fig. 9-24. In this circuit:

l_1 and l_2 are the leakage inductances of the two windings.

r_1 and r_2 are the winding resistances (nearly equal to the dc resistances at low frequency).

M is the mutual inductance (which is nonlinear in iron-core devices).

R represents eddy-current and hysteresis loss and is a function of frequency and level.

TABLE 9-3 Basic Bridge Configurations

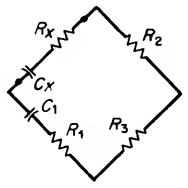
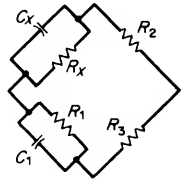
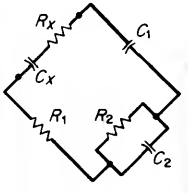
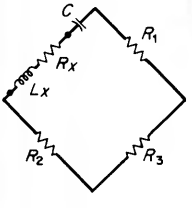
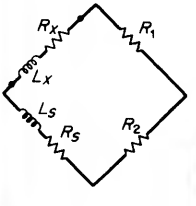
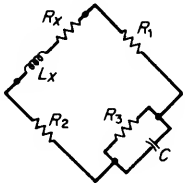
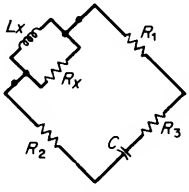
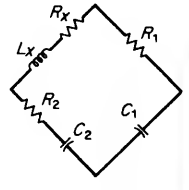
Series Capacitance Comparison Bridge Most common capacitance bridge circuit. Usually R_3 and R_1 are variable so that bridge reads capacitance and D.		$C_x = C_1 \frac{R_2}{R_3}$ $R_x = R_1 \frac{R_2}{R_3}$ $D = \omega C_1 R_1$ Series components
Parallel Capacitance Comparison Bridge Used for parallel capacitance measurements and for high D measurements where R_1 of the series bridge would be impractically high.		$C_x = C_1 \frac{R_2}{R_3}$ $R_x = R_1 \frac{R_2}{R_3}$ $D = \frac{1}{\omega C_1 R_1}$ Parallel components
Schering Bridge Used in high-voltage bridges with a high-voltage capacitor as C_1. Used in high-frequency bridges because variable capacitors can be used for both adjustments. Used in dielectric measurements because C_2 gives high-resolution loss measurement. A series combination of C_2 and R_2 reads parallel capacitance.		$C_x = C_1 \frac{R_2}{R_1}$ $R_x = R_1 \frac{C_2}{C_1}$ $D_x = \omega R_2 C_2$ Series components
Series Resonant Bridge Rarely used to measure L because of frequency dependence. Parallel version has some equations for parallel components, and is used with variable frequency source to measure resonant frequency of coils.		$L_x = \frac{1}{\omega^2 C}$ $R_x = \frac{R_1 R_2}{R_3}$ Series components
Maxwell Inductance Comparison Bridge Basic inductance comparison bridge. Additional resistance added to inductor with lower Q to get resistance balance. Series circuit shown; could also be parallel.		$L_x = L_S \frac{R_1}{R_2}$ $R_x = R_S \frac{R_1}{R_2}$ Series components

TABLE 9-3 (Continued)

Maxwell-Wien Bridge Most common inductance bridge. If R_2 and R_3 are variable, it reads L and Q. If R_3 and C are variable, it reads L and R.		$L_x = R_1R_2C$ $R_x = \frac{R_1R_2}{R_3}$ $Q = \omega CR_3$ Series components
Hay Bridge Useful for parallel inductance measurements or high-Q measurements where R_3 of Maxwell bridge would have to be impractically high. Also used for measurements with dc current bias because all current applied across bridge flows in L_x.		$L_x = R_1R_2C$ $R_x = \frac{R_1R_2}{R_3}$ $Q = \frac{1}{\omega R_3C}$ Parallel components
Series Owen Bridge Used in precision inductance bridges where R_2 is a high-resolution decade reading L_x, and C_2 reads G_x. Parallel Owen bridge, reading parallel L_x and G_x, is formed if R_2 and C_2 connect in parallel.		$L_x = R_1R_2C_1$ $G_x = \frac{C_2}{R_1C_1}$ Series components

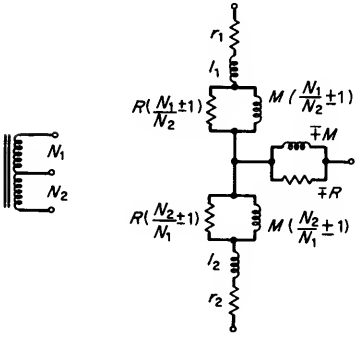


FIG 9-24 An equivalent circuit for a transformer.

N_1 and N_2 are the numbers of turns.

The signs depend on the connection of the windings, the upper one representing an "aiding" connection as would be used in ratio arms.

Capacitance is not shown. Although distributed, it can be approximated by lumped capacitance between the terminals at low frequencies and thus be considered an external load. One should note that even above resonance, a tapped coil still maintains its attractive features.

Ratio Accuracy. If a voltage is applied across the pair of windings in Fig. 9-24, the ratio of the two voltages is very nearly

$$\frac{E_1}{E_2} = \frac{N_1}{N_2}\left[1 + \left(\frac{N_2(r_1 + j\omega l_1) - N_1(r_2 + j\omega l_2)}{N_1 + N_2}\right)\left(\frac{1}{R} + \frac{1}{j\omega M}\right)\right] \tag{9-5-4}$$

This ratio can easily have an error of less than 1 ppm even if the ratio is quite large. To obtain an accurate ratio, one should:

1. Use a high-μ, low-loss core so that M and R are large. Tape wound, toroidal cores with a μ of over 100,000 are available. The material should be very thin to reduce eddy-current losses.
2. Intermix the windings to get low leakage inductance.
3. Use as many turns as possible to get the l/M ratio small, for l/M is approximately proportional to $1/N$.
4. Use as large wire as possible and use the same wire size for both windings so that $N_1 r_2 = N_2 r_1$.

Not only is such a ratio accurate; it is also extremely constant. Even a large change in μ from shock or overload makes only a small change in ratio which often can be removed by cyclic demagnetization.

Loading Error. If the N_2 winding of an inductively coupled voltage divider is shunted by a load Z_y, the resulting error in ratio is approximately

$$\frac{(N_1/N_2)(r_2 + j\omega l_2) + (N_2/N_1)(r_1 + j\omega l_1)}{[(N_1 + N_2)/N_2]Z_y}\,100\%$$

If $N_1 = N_2$, this is very nearly equal to the loading error that would result if the ratio arms consisted of winding resistance and leakage inductance only. Thus, inductively coupled ratio arms are very accommodating: They exhibit a very high value of M and R to provide an excellent ratio and to present a high impedance to the supply, but they exhibit a very low effective impedance to a shunt load! Such behavior is, of course, due to their coupling and the resulting reflected impedance.

Ratio accuracy and immunity to loading are the reasons that "transformer" ratio arms are used in three-terminal capacitance bridges, such as the one shown in Fig. 9-25. Here, one stray capacitance shunts the detector, while the other shunts one winding. Bridges are made where

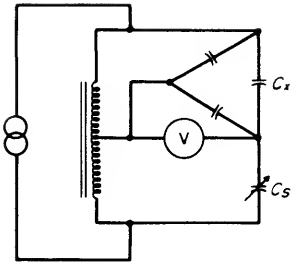


FIG 9-25 A three-terminal capacitance bridge with inductively coupled ratio arms.

either stray capacitance could be as high as 1 μF at 1 kHz and the resulting error would be less than 0.01 percent [27].

Transformer Connections. Three basic connections of a transformer are shown in Fig. 9-26. Each has certain advantages of accuracy under certain conditions, as described below.

Voltage-divider connection [23] (Fig. 9-26*a*). This connection has the low errors described above and is preferred for high-accuracy measurements on low impedances because the impedances of the other two bridge arms do not affect the transformer ratio. Usually a second transformer is used to drive such a bridge to provide isolation.

Third-winding drive [23] (Fig. 9-26*b*). This circuit is particularly good when comparing high impedances because the open-circuit voltage ratio of the two secondary windings depends only on the ratio of the mutual inductances between these windings and the primary, which can be very accurate, and not on winding resistance and leakage inductance. However, here these residual winding impedances are part of the other bridge arms and, thus, cause large errors when low impedances are measured.

Lynch connection [24] (Fig. 9-26*c*). In this autotransformer connection, full input voltage is applied across one-half of the transformer, and as a result, loading across this winding has no effect on the balance condition. Thus, this connection is useful when extreme loading may be present such as in *in situ* [25] measurements. However, this connection has the poorest

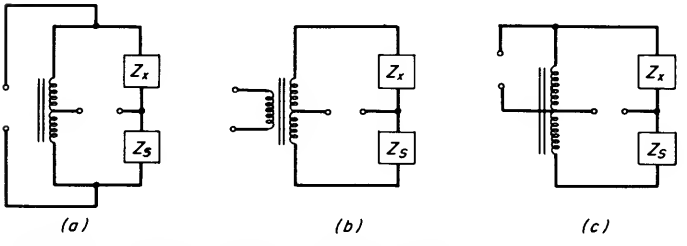


FIG 9-26 Methods of driving transformer ratio arms.

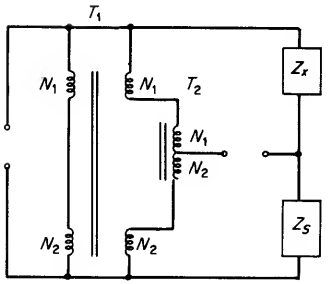


FIG 9-27 A compound transformer connection.

open-circuit voltage ratio accuracy because of its nonsymmetry. Also, the winding impedance of the undriven winding and the reflected winding impedances of the driven winding are both effectively in series with Z_S.

An interesting *compound connection* [26] (Fig. 9-27) combines the advantages of the first two circuits. Here very little current flows in the secondary of T_1 or in T_2 because the voltage induced in T_1 from its primary is nearly equal to the voltage applied to the series connection of the other windings. Because only a small current flows, the voltage ratio is quite independent of winding resistance and leakage inductance.

What is shown in Fig. 9-27 as two transformers can be made as one unit by using two toroidal cores. The primary of T_1 is wound on one core, but the secondaries of T_1 and the T_2 windings are formed with windings that enclose both cores.

Transformers are not very linear devices, and therefore one would not expect that the source and detector of a bridge using them could be reversed without changing the balance conditions. However, as long as saturation is avoided, the difference caused by this reversal is very small because the dependence of the ratio on the nonlinear quantities, M and R in Fig. 9-24, is so small. The design requirements are quite different, depending on the location of the transformer. If used in the input, it must have adequate turns and a proper core to accept the applied voltage without saturation. If used in the output, it should be shielded to avoid magnetic pickup. Several bridges use transformers in the input and in the output, Fig. 9-28, to obtain high ratios, to apply equal power to the standard and unknown, or to provide an impedance match to the detector (see Fig. 9-27).

Practical Transformer-ratio-arm Bridges. While the advantages of this type of bridge are very important, there are the two following disadvantages that make their design more complex than conventional four-arm bridges:

1. The inductively coupled arms use up two arms of the bridge, which leaves one arm for the unknown and only one arm for both adjustments

(real and imaginary). The transformer ratio may be easily switched to change ranges, but a high-resolution variable voltage, as required for a balancing adjustment, is difficult.

2. The inverting type (refer to Sec. 9-5-1) of bridge that compares an unknown inductance with a capacitance is impossible if the bridge is limited to basic four-arm circuits.

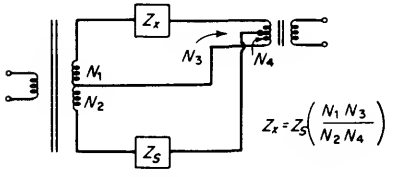


FIG 9-28 A bridge with a ratio transformer in the input and the output.

Both these limitations may be overcome by taking advantage of the excellent immunity to loading, which allows the use of T networks whose direct impedance gives the desired function of variable elements. Several practical bridge circuits are illustrated in Table 9-4. The YΔ transformation (Sec. 9-1) is helpful in obtaining the balance equation of two of them.

9-5-4 Special-purpose Bridges

Hague [16] discusses a wide variety of bridges designed for special purposes. Many of these are simply modifications in the method of adjustment, and even though some have additional network branches, they can be reduced to one of the common four-arm bridges by use of the YΔ transformation. Another catalog of these bridges would be of little use, but a mention of the basic problems is worthwhile, along with some of the more interesting solutions.

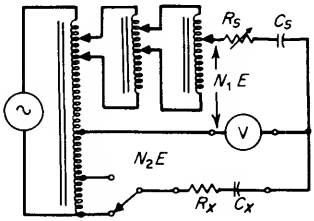
High-impedance Measurements. The techniques used for high-resistance dc measurements (Sec. 9-4-3) are applicable to ac measurements, particularly the techniques of making three-terminal measurements. Measurements on small capacitors are subject to large errors owing to the shunting of the unknown by stray capacitance unless a guarded three-terminal configuration is used. Most commercial *CRL* bridges use capacitance bridge circuits which allow substantial capacitance to a third guard terminal simply because the standard capacitor is so large that a small shunting stray capacitance has little effect. Likewise, Wagner guard circuits for precision capacitance measurements have been available [28]. However, the use of transformer-ratio-arm bridges in the last

TABLE 9-4 Bridges with Inductively Coupled Ratio Arms

Decade-voltage Adjustment (ESI† 701)

A decade transformer is used to provide a high-resolution voltage adjustment as the main balance. N_2 is switched to change range. An additional transformer provides a percent deviation reading.

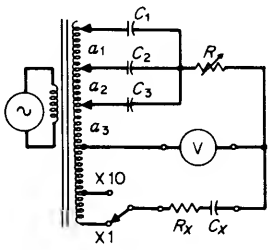
$$C_x = C_S \frac{N_1}{N_2} \qquad D = \omega R_S C_S$$



Digital-summing Adjustment (GR‡ 1615)

The currents from six capacitors, each with a step variable voltage applied, are added to give a high-resolution adjustment (6 digits).

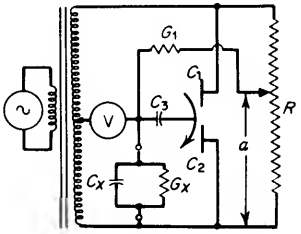
$$C_x = a_1C_1 + a_2C_2 + a_3C_3 + \cdots$$
$$C_1 = 10C_2 = 100C_3 \quad \text{etc.}$$
$$D_x = \omega R(C_1 + C_2 + C_3 + \cdots)$$



Differential-capacitor Adjustment (BEC¶ 75C)

A differential capacitor provides a variable voltage with constant output capacitance if $C_1 + C_2$ is constant. C_3 reduces range. The conductance balance shown is used on many bridges reading parallel G.

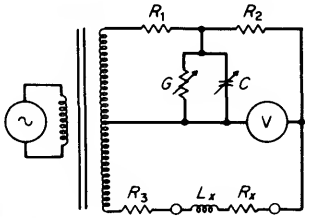
$$C_x = \frac{(C_1 - C_2)C_3}{C_1 + C_2 + C_3}$$
$$G_x = \frac{G_1(2a - 1)}{1 + G_1R(a)(1 - a)}$$



"Opposed T" Inductance Bridge

YΔ transformation of T network gives simulated inductance. "Spoiler" R_3 is necessary for balance unless $R_x > R_1 + R_2$.

$$L_x = R_1R_2C$$
$$R_x = R_1R_2G \qquad \text{if } R_3 = R_1 + R_2$$



† Electro-Scientific Industries, Inc.
‡ General Radio Company.
¶ Boonton Electronics Corporation.

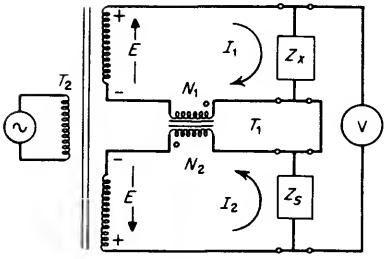


FIG 9-29 A four-terminal bridge with a transformer used to provide an accurate current ratio.

decade has made this measurement very convenient and has almost completely replaced the use of guard circuits which require additional balance adjustments.

Low-impedance Measurements. Likewise, the techniques used for low-resistance dc measurements may be adapted for low-impedance ac measurements. Several commercial bridges make four-terminal measurements on large capacitors simply by placing the lead resistances in the adjacent arms as illustrated in Fig. 9-18. Also, the Kelvin double bridge has been adapted for ac measurements [26, 29] even though it may require two adjustments in both the main and auxiliary bridge arms because these arms are complex.

The use of transformers makes possible four-terminal ac bridges that have no dc equivalents and are somewhat analogous (dual) to the transformer type of three-terminal bridges described above. Just as a transformer can produce voltages of precise ratio, it can also produce currents of precise ratio. In Fig. 9-29 transformer T_1 has no flux if $N_1I_1 = N_2I_2$ and, therefore, presents low impedances to these currents. However, any unbalance current is presented the full open-circuit inductance and thus is kept very low. The other transformer T_2 supplies isolated voltages which make the four-terminal connection possible but which need not be precise.

Another application of a transformer is to use it to inject a voltage into one branch of a bridge to compensate for an undesired voltage in another [30]. In the four-terminal capacitance bridge of Fig. 9-30, the voltage

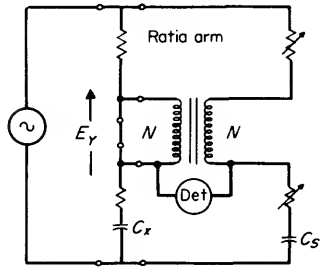


FIG 9-30 A four-terminal capacitance bridge (General Radio Company 1617).

drop E_Y, caused by lead and yoke resistance, is injected into the opposite side of the bridge. This works very well only if one side of the bridge has a much higher impedance than the other [31].

Mutual-inductance Measurements. Mutual inductance produces a *transfer* impedance, actually just one component of a transfer impedance that in general has a resistive (loss) component (see Sec. 9-1). There is some disagreement on whether a series or parallel representation of this complex impedance should be used. The parallel representation is used in Fig. 9-24 because it simulates circuit behavior better. Here mutual inductance becomes the effective parallel inductance of the open-circuit transfer impedance between a pair of coupled coils.

This mutual inductance produces a special kind of four-terminal impedance, for if Fig. 9-24 is compared with Fig. 9-2, one sees that:

1. Because the coils may be tied together at one end, there are only two "lead" impedances, and the network forms a T instead of an H.
2. These lead impedances may be very large, larger than the mutual impedance itself, unless the turns ratio is unity.
3. The mutual impedance may be positive or negative depending on the connection.

Four-terminal bridge techniques may be used to measure mutual inductance. The Carey-Foster bridge [32] and several others put the T arms (lead impedances) in the source and in an adjacent arm, as is done in the dc bridge of Fig. 9-18. The balance equation for these bridges is very complicated if the mutual impedance has loss. A pure mutual inductance can be nulled with a capacitor by using Campbell's frequency bridge [33] (Fig. 9-31). The circuit of Fig. 9-32 will measure the parallel mutual components if the connection giving negative quantities is used and the Q is less than $1/\omega C(R_1 + R_2)$.

In spite of the many interesting circuits possible, mutual impedance is most often measured by a series of two-terminal measurements. From

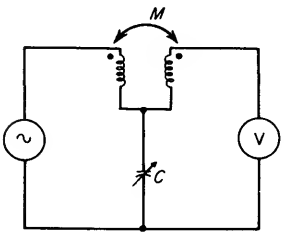


FIG 9-31 Campbell's frequency bridge used to measure lossless mutual inductance.

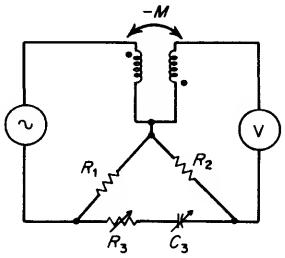


FIG 9-32 A null network for measuring lossy mutual inductance.

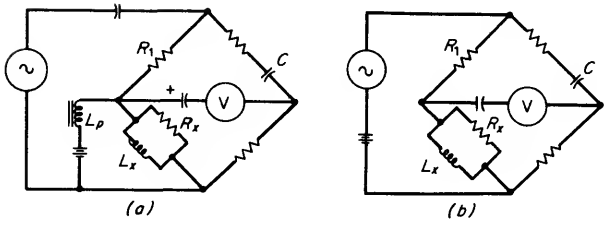


FIG 9-33 Methods of supplying dc current to make incremental inductance measurements. (a) Parallel feed, (b) series feed.

Fig. 9-24 one can see that if the two coils are measured in series, first aiding and then opposing, the difference in impedance is four times the mutual impedance (a series-to-parallel conversion is necessary to get the parallel quantities shown).

Applying Bias Voltage or Current. The capacitance of many capacitors changes as dc voltage is applied. The capacitance of semiconductor devices changes drastically. Likewise, the inductance and loss of iron-core inductors changes with applied dc current, and resistive devices change value with applied voltage (varistors) or heat (thermistors). The slope of the impedance curve at some dc operating point may be measured with a small ac signal and is called the *incremental* impedance. A bridge circuit is convenient for this purpose, but the bridge circuit must be adapted to make such measurements without error or damage.

Two basic methods of introducing bias are used: series and parallel feed. Figure 9-33 illustrates circuits for measuring incremental inductance. With series feed no error is introduced in the bridge circuit, but the current is limited by the power rating of the bridge resistor in series, R_1. There is no such limit with parallel feed, but the bridge will measure the parallel combination of the unknown and the impedance of the feed circuit so that a correction must be made. Note that the Hay bridge is often used for this purpose because its capacitor blocks dc current from flowing in the other bridge arms. Additional blocking capacitors are added in the source and detector as required.

A Schering or series comparison bridge is generally used for biased capacitor measurements to provide dc blocking. The voltage is usually applied by series feed such that all the dc voltage appears across the unknown capacitor.

Active Bridges. The reliability of semiconductor devices together with the precision of feedback amplifiers has made it practical to include active elements in bridge circuits. Their use provides additional flexibility in design mainly by their ability to provide isolation and phase inversion. Unity gain amplifiers, both inverting and noninverting, with

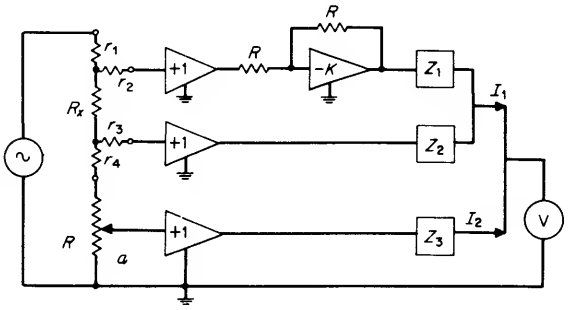


FIG 9-34 Logan's four-terminal ac resistance bridge.

long-term stability of better than 0.01 percent, are possible at low frequencies. The most critical part of the inverting amplifier can be the passive components used for current summing.

An example of an active bridge is Logan's bridge [34] (Fig. 9-34) which was designed for ac resistivity measurements using a four-point probe. Impedances Z_1 and Z_2 and the amplifiers driving them form a differential amplifier whose CMR (common-mode rejection) makes the current I_1 independent of R_S and the lead resistance r_4.

Another common use of active devices is the active guard circuit of Fig. 9-35 used for three-terminal measurements. Here C_a has no voltage applied to it and thus passes no current if the gain of the amplifier is unity, and C_b loads the amplifier and not the bridge; therefore no error is introduced.

9-5-5 *Automatic and Semiautomatic Bridges*

For production or inspection measurements the manual process of balancing a bridge is too slow and therefore too expensive to be practical. While impedance meters are often used for less accurate measurements, and comparators are used when many similar components are to be tested and suitable standards are available, for many applications the

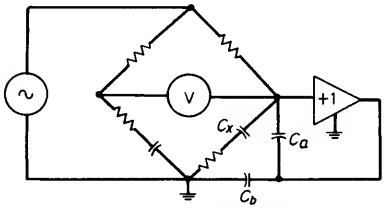


FIG 9-35 An active guard circuit.

combination of range, accuracy, and direct readout of the bridge is required.

Mechanically Balanced Bridges. Several commercial automatic bridges have been made by connecting servo motors to the shafts of the variable bridge components and driving them with the amplified bridge output. Phase-sensitive detectors are necessary to distinguish the sense of the unbalance voltage and to separate the components of voltage resulting from each of the two adjustments. In general, the two balance adjustments of an ac bridge do not produce orthogonal unbalance voltages, and as a result there is apt to be interaction between the two adjustments, which in manual bridges is called a *sliding null* [35]. Fortunately, as long as the phase angle of impedance is not too large and the adjustments are continuously variable, the interaction only increases the time necessary for balance.

Electrically Balanced Semiautomatic Bridges. One interesting commercial CRL bridge [36] uses the nonlinearity of biased diodes to make an automatic loss (D or Q) balance and a phase-sensitive detector to facilitate the balance of the main adjustment, which is manual. As a result, the value of a component may be measured very quickly even if it has a phase angle that would usually result in a sliding null. The automatic loss balance has no readout. However, if the D or Q value is desired, a manual balance may be made by adjustment of the loss balance alone after the main balance is made in the semiautomatic mode.

Digitally Balanced Automatic Bridges. A fully automatic, electronically balanced capacitance bridge has been made that uses a digital servo technique for both balances [37]. Transistor switches are used to adjust the attenuation of conductance, summing-type voltage dividers in an active bridge network. The switches are controlled by phase-sensitive detectors on the bridge output, and their settings provide information for a digital readout.

9-6 Radio-frequency Impedance Measurements

9-6-1 Problems at Radio Frequency

As frequency increases, many of the residual parameters in the equivalent circuits for the various types of circuit components (see Sec. 9-2) cause increasingly larger errors. Wire-wound rheostats, often used in lower-accuracy bridges, become almost useless above 100 kHz. The effective value of large capacitors changes appreciably because of lead inductance; capacitors larger than 1,000 pF are to be avoided above 1 MHz. Low-frequency iron-core inductors are useless at radio frequency, although coils designed for a particular frequency range are

useful up to quite high frequencies. The impedance of bridge wiring and switching is added to the residual parameters of the components themselves. Distributed capacitance effects become important particularly for large resistances, and mutual inductance between leads must be considered.

As a result of these considerations, rf bridges usually have variable capacitors as adjustable components and only small fixed resistors. Range switching is limited, shielding becomes complex, and wiring is carefully laid out to avoid large coupling loops. Very high or very low impedances are avoided so that the bridge range becomes limited. This trend continues until at ultrahigh frequency all wiring is coaxial, discrete components almost completely disappear, and only impedances near the characteristic impedance of the system can be measured with any accuracy (see Chap. 17).

9-6-2 *Radio-frequency Bridges*

General Radio Company RF Bridge (GR 1606) [38]. A favorite for use at radio frequencies is the Schering bridge, because both null conditions may be met with the use of only variable capacitors. An example of a commercial instrument with this circuit is the GR 1606 of Fig. 9-36. This is a substitution bridge, for one adjustment C_p is in series with the unknown. Both components of the unknown are read as difference measurements. Using subscripts 1 and 2 to refer to successive balances made with the terminals shorted and then with the unknown in place, we have

$$R_x = \frac{R_B}{C_N}(C_{A2} - C_{A1})$$
$$X_x = \frac{1}{\omega}\left(\frac{1}{C_{p2}} - \frac{1}{C_{p1}}\right) \tag{9-6-1}$$

The dials associated with both adjustments are calibrated in ohms, but the reactance dial must be divided by the frequency in megahertz to

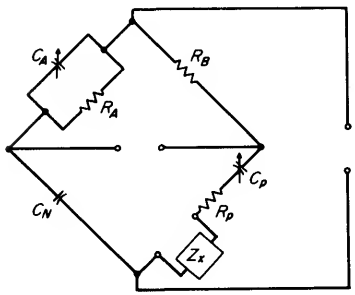


FIG 9-36 A series substitution Schering bridge (General Radio Company 1606).

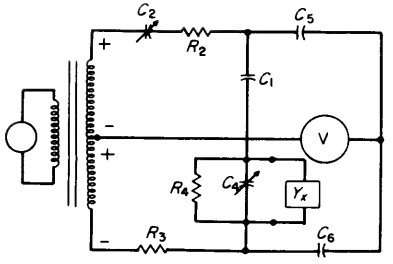


FIG 9-37 The *RX* meter (Hewlett-Packard Company 250B).

obtain the correct result. This instrument covers the 400-kHz to 60-MHz range.

Hewlett-Packard Company *RX* Meter (HP 250-B) [39]. This instrument is a parallel substitution device measuring admittance. It is shown in Fig. 9-37 as a transformer bridge with opposed T networks. The four main arms are similar to those of a Schering bridge while the small capacitors C_5 and C_6 are used as a summing connection to the detector. This instrument includes an internal oscillator and detector and operates from 500 kHz to 230 MHz.

Wayne Kerr Bridges (B601, B801, and B901) [40]. This series of instruments uses the transformer-ratio-arm principle well up into the very high frequency range. The transformers used have a somewhat novel structure and have very few turns to keep the leakage inductances low. Fixed, switched resistors are used for the conductance balance to avoid the large frequency errors in rheostats.

Admittance Meter (GR 1602) [41]. This novel device, illustrated in Fig. 9-38, is classified here as a bridge because it is a nulled device which compares passive standards with an unknown. Mutual inductances M_B, M_G, and M_x are varied by rotation to inject voltages into the detector proportional to the currents in the susceptance standard B_S, the conductance standard G_S, and the unknown admittance $G_x + jB_x$. Because the same voltage is applied to all three impedances, the detector is nulled when

$$M_x(G_x + jB_x) = M_G G_S + M_B j B_S \tag{9-6-2}$$

The dials associated with M_G and M_D are calibrated in terms of G_x and B_x, and M_x is used as a multiplier to extend the range. This instrument is direct reading over a 40- to 1,500-MHz range.

9-6-3 *T Networks* [42]

While any network capable of giving a null output could be considered a bridge in a broad sense, there is a large class of null networks with a com-

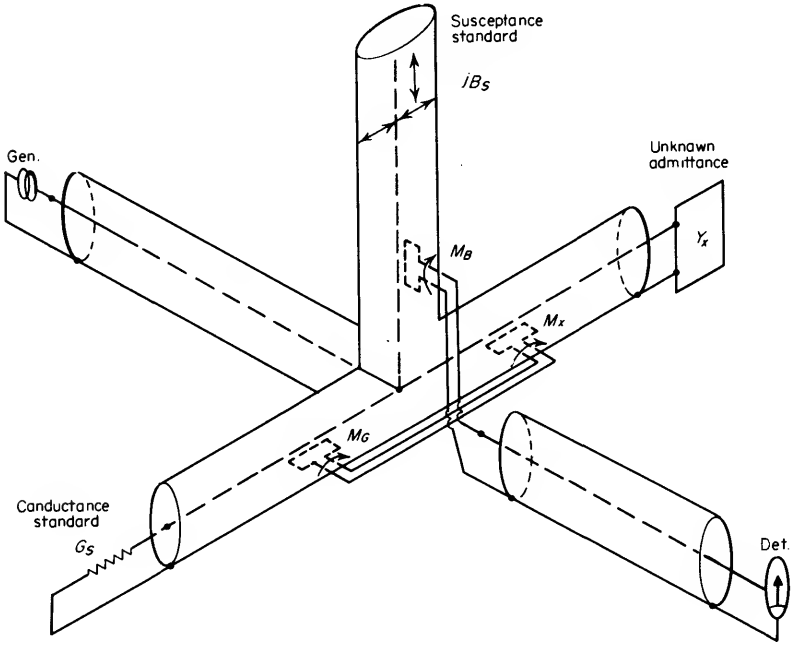


FIG 9-38 An admittance meter (General Radio Company 1602).

mon input and output terminal (without the use of transformers) that are based on the simple T configuration and, therefore, classified as T networks. While these networks have been used at low frequencies, their main application is at radio frequencies where their common input-output terminal is very important to avoid unwanted coupling.

The null conditions for these networks are easily obtained by setting at zero the sum of all the short-circuit transfer admittances of all network paths from source to detector. If there are only two paths, the sum of the short-circuit transfer impedances is also zero. For a single T network (Fig. 9-4) this transfer impedance is $Z_1 + Z_3 + Z_1Z_3/Z_2$. It is the last term of this expression that makes a null possible, for it can be used to form an effective inductor from a capacitor and two resistors or a negative resistor from two capacitors and one resistor. All T null networks are frequency dependent in both balances, for without a phase inversion, inductance or simulated inductance must balance a capacitance, and simulated negative resistors, which always contain $-\omega^2$ or $-1/\omega^2$ factors, must be used to balance resistors.

The Bridged T. If a single T is shunted by an impedance Z_4, a *bridged* or

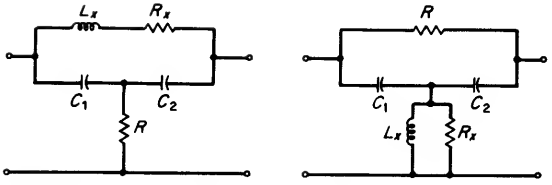


FIG 9-39 Two bridged-T null circuits.

shunted T is established. The balance expression is

$$Z_1 + Z_3 + \frac{Z_1 Z_3}{Z_4} + Z_4 = 0 \tag{9-6-3}$$

There is some similarity to the simple bridge equation, the difference being the addition of the first two terms and a very important sign. Such a circuit can give a complete null only if it contains at least one inductor. The two simplest circuits are shown in Fig. 9-39. While such networks are often used in filters, they are rarely used for impedance measurements.

Twin-T Networks. Two T networks in parallel will null if the transfer impedance of one is the negative of the other. Many twin-T networks are possible; the most important one for impedance measurements is shown in Fig. 9-40. This is not the same twin T used in many oscillators and filters. Its main advantage is that the two grounded variable capacitors can be used for the two necessary adjustments. The balance equations are

$$C_1 + C_2 + C_3 + \frac{C_1 C_3}{C_4}(1 + R_6 G_5) = \frac{1}{\omega^2 L_2} \tag{9-6-4}$$

and

$$\omega^2 R_6 C_1 C_3 \left(1 + \frac{C_5}{C_4}\right) = G_2$$

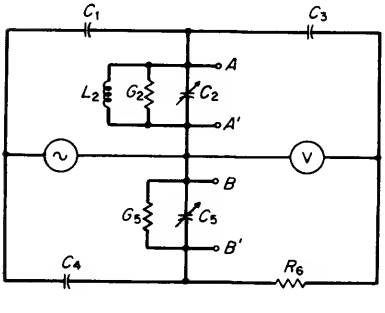


FIG 9-40 The twin-T admittance-measuring circuit.

from which

$$B_{xA} = \omega \, \Delta C_{2A} \qquad \text{or} \qquad C_{xA} = \Delta C_{2A}$$

$$G_{xA} = \frac{\omega^2 C_1 C_3}{C_4} R_6 \, \Delta C_{5A}$$

where ΔC_{2A} and ΔC_{5A} are the changes in the settings of these two capacitors when the unknown is added to terminals A and A'.

A commercial instrument [43] was made once using this circuit (except that $G_5 = 0$). It was useful for measuring high impedances, particularly measurements on dielectric samples, over a range from 460 kHz to 40 MHz.

The same basic network has been used for precision measurements [44] with additional terminals B and B' across C_5. With the unknown connected to these B terminals, the admittance of the unknown is

$$B_{xB} = \omega \, \Delta C_{5B} \qquad \text{or} \qquad C_{xB} = \Delta C_{5B}$$

$$G_{xB} = \frac{C_4 \, \Delta C_{2B}}{C_1 C_3 R_6} \tag{9-6-5}$$

where ΔC_{2B} and ΔC_{5B} are the changes in these settings when the unknown is added to the B terminals. If a given unknown is measured first on the A terminals and then on the B terminals, $G_{xA} = G_{xB} = G_x$, which may be determined from the expression

$$G_x^2 = \omega^2 \, \Delta C_{5A} \, \Delta C_{2B} \tag{9-6-6}$$

Thus, both B_x and G_x can be calculated from changes in capacitance without knowing the exact value of the other network impedances. This is probably the most accurate method of measuring resistance in the rf range.

9-6-4 *Resonance Methods*

A resonance, as well as a null, is a repeatable condition in which the circuit elements can have a defined relationship. Parallel resonance is best suited for high-impedance measurements and series resonance for low-impedance measurements.

Parallel-resonance Methods. If the capacitor in the parallel resonant circuit of Fig. 9-41 is adjusted for a resonance (peak reading on the voltmeter) with the unknown terminals open-circuited to give a reading of C_1 and then with an unknown admittance connected to give a reading of C_2, the susceptance of the unknown B_x is $\omega(C_1 - C_2)$. This simple principle has been used in an rf capacitance meter. Note that the reading is independent of the values of the other circuit elements.

The conductance of the unknown may be obtained by making a third

resonance with a known conductance G_S connected and noting all three peak voltmeter readings. Now

$$G_x = G_S \frac{V_3(V_1 - V_2)}{V_2(V_1 - V_3)} \tag{9-6-7}$$

This is what is called the *conductance variation method.*

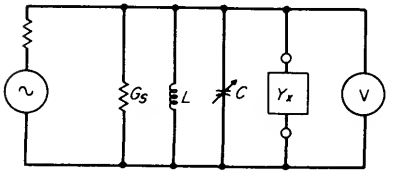


FIG 9-41 Parallel-resonance measurement circuit.

An alternative way to obtain G_x, called *susceptance variation,* is to detune the capacitor until the voltage indicator is 0.707 of the peak voltage by both addition and removal of capacitance. If the higher capacitance reading is C' and the lower is C'', then $G = \omega(C' - C'')/2$. Here G is the total circuit conductance. The unknown is determined by making this measurement with the unknown first in place and then removed and by calculating the difference in conductance.

This susceptance variation method has been used for precise dielectric loss measurements [45]. In one method [46] a negative conductance, made from an active circuit, was used to cancel out most of the loss of the

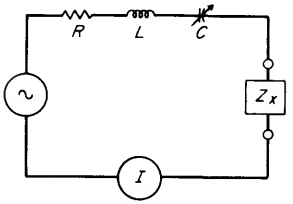


FIG 9-42 Series-resonance measurement circuit.

inductor and so to make a very high Q circuit. This gave dissipation-factor resolution of a few parts per million.

Series-resonance Methods. The duals of the above methods for the circuit of Fig. 9-42 are called the *resistance* and *reactance* variation methods.

They give the effective series resistance and reactance of the unknown. With this circuit the initial measurements are made with the unknown terminals shorted; the reactance is proportional to the difference in elastance (reciprocal capacitance), and the resistance may be determined by meter readings or by an elastance difference.

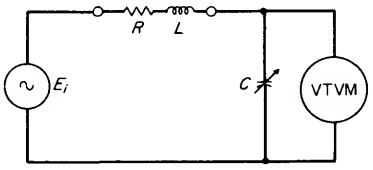


FIG 9-43 The Q meter.

The Q Meter (Resonant-rise Method) [47]. The best-known resonant method is used for the Q meter shown in Fig. 9-43. If the capacitor is lossless, the voltage across it is

$$E_o = E \frac{1/j\omega C}{R_x + j(\omega L_x - 1/\omega C)} \tag{9-6-8}$$

At resonance, $\omega L_x = 1/\omega C$, and the magnitude of the output voltage is

$$\frac{E_o}{E_i} = \frac{1}{\omega R_x C} = \frac{\omega L_x}{R_x} = Q_x \tag{9-6-9}$$

The meter of this instrument can be calibrated in terms of Q, and the variable-capacitor reading can be used to calculate inductance.

9-6-5 *The RF Meter Methods*

The basic meter methods of Sec. 9-3 may be used at radio frequencies with appropriate meters. The comparison method of measuring voltages across known and unknown impedances is particularly useful at radio frequencies now that low-level voltmeters are available.

The RF Vector Impedance Meter (H-P 4815) [14]. This instrument is somewhat similar to the lower-frequency meter (Sec. 9-3) but uses a phase-locked sampling system to obtain and operate in a frequency range of from 0.5 to 108 MHz. As shown in Fig. 9-44, an rf current is supplied to the unknown, measured by a toroidal pickup coil, sampled, amplified, detected, and applied to a modulator to form an automatic current-level control. The voltage is similarly sampled and, because the current is held constant, is a measure of impedance magnitude. A phase detector measures θ.

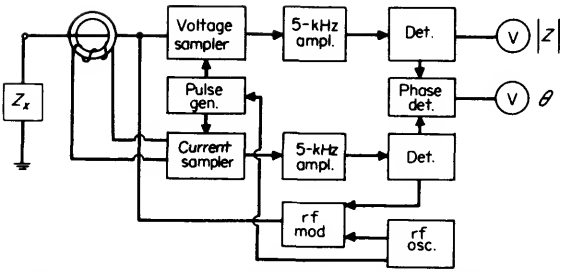


FIG 9-44 The rf vector impedance meter (Hewlett-Packard Company 4815).

9-7 Precision Measurements

9-7-1 Standardization of Impedance Units

Determination of the Ohm[1] [48, 49]. In this country the National Bureau of Standards (NBS) is responsible for the determination, maintenance, and dissemination of standards for all units of measure, including the electrical unit of impedance—the ohm—and the related units of capacitance and inductance. In 1864, long before NBS existed, a British committee led by James Clerk Maxwell made the first determination of the ohm. They gave values to certain wire resistors, based on the mechanical system of units and Wilhelm E. Weber's proposed electromagnetic system of units. While several of these resistors were distributed to various laboratories, a lack of standardization facilities and poor transportation made a need for a standard that could be reproduced independently. This resulted in the *mercury ohm*, a column of mercury of given dimensions measured at a given temperature.

Soon after NBS was established in 1901, it obtained four resistors made in Germany and certified by the Reichsanstalt, the oldest national standards laboratory. The mean value of these resistors and others calibrated from them became our national standard. They were replaced by a sealed NBS-Rosa type of 1-Ω resistor in 1909 and by the present Thomas type of resistor in 1932, as these proved more stable in international comparisons. During this period an international committee determined the *international ohm* on the basis of the values maintained by several national laboratories.

Meanwhile, experimenters were redetermining the ohm from the basic units by formulas relating the dimensions of a coil to its inductance. This was a painstaking procedure, for the physical measurements had to be extremely precise if accuracy of a few parts per million were to be

[1] See also Chap. 1.

realized. The ohm was determined from the calculated inductance by several measurement circuits, which introduced the unit of time that relates inductance (and capacitance) to resistance. The determined value differed from the international values, and in 1948 the *absolute ohm* came into use by international agreement. It was 495 ppm smaller than the mean international ohm, and hence values in the new units were 495 ppm larger.

Since then, additional determinations have been made with the use of calculable inductors and, more recently, the Thompson-Lampard calculable capacitor [50], which is more easily and more precisely measured. A cross section of such a capacitor made from four solid metal rods is shown in Fig. 9-45. It can be proved that if there is symmetry about one axis, the direct capacitance between either opposing pair of rods, with the other pair guarded, will give a capacitance in picofarads of $10^{19} \times L(\ln 2)/4\pi^2c^2$, where L is in meters and c is the speed of light in meters per second. Moreover, small errors due to nonsymmetry can be removed to a very large degree by taking the mean value of two measurements with the use of first one opposing pair and then the other.

Unfortunately, the capacitance so determined is quite small, approximately 1.95 pF/m. Precise transformer-ratio-arm bridges are used to transfer the measurement up to a more reasonable capacitance value. Then a special *quad* (for quadrature) bridge compares capacitance and frequency with an ac resistance. The value is then transferred to a resistor whose ac/dc difference is calculable, and then by dc-ratio methods to the 1-Ω value. Cutkosky [51] of NBS made this determination and came within 2.3 ppm of the 1-Ω value maintained by NBS and had an estimated uncertainty of 2.1 ppm. More recently Thompson [52] made a similar determination with an estimated uncertainty of 0.2 ppm, but with an accuracy limited to ±0.7 ppm because of the uncertainty of the speed of light.

National Bureau of Standards Calibrations. Ohm determinations are

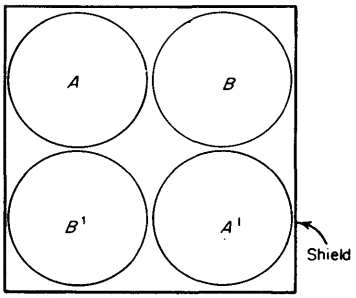


FIG 9-45 Cross section of a Thompson-Lampard calculable capacitor.

made by NBS only every few years, and the value is maintained by a reference set of 10 Thomas 1-Ω resistors. These are intercompared regularly with a working set that is used for actual intercomparisons with standards sent in by customers. A similar procedure is used for capacitance and inductance calibrations. The NBS will calibrate over a wide range of values, but highest accuracy is given at medium values. They will now state 0.08-ppm accuracy for resistance at 1 Ω and 10 kΩ, 5 ppm for three-terminal capacitance, and 200 ppm for inductors of medium value. A complete list of the available calibrations is given in NBS Miscellaneous Publication 250, which is revised regularly.

National Bureau of Standards is attempting to limit its calibrations to the basic fixed standard for each unit and trying to get other laboratories to extend the calibration to lesser standards, variable components, and instruments by proved techniques. In this way they can put more of their effort into new and improved basic standards, which are urgently needed.

Traceability. The word *traceability* is used to describe the chain of calibrations that can link the assigned value of any component to the national standard. Evidence of each step in this chain is required by the National Aeronautics and Space Administration (NASA) and military standards [53] in an attempt to ensure that suppliers of electrical components are making accurate measurements. While traceability may be necessary to meet this goal, it is not sufficient to ensure high accuracy, for a standard may change in value between the times it is measured and used, and good measurement technique is essential, especially when measuring odd values at odd frequencies. In spite of these and other inherent limitations, the concept of traceability has greatly improved the general level of measurement accuracy, particularly at the component-inspection level.

Associated with traceability are two additional terms: *recall cycle* and *accuracy ratio*. The first represents the time period between calibrations which is left indefinite by some standards. The United States Navy [54] however, has established recommended recall cycles for a large variety of commercial standards and instruments. Accuracy ratio is the ratio of the accuracy of the device being tested to the accuracy of the standard used for calibration. While a high accuracy ratio results in a short traceability chain, in many cases it is not feasible, particularly at higher accuracy levels. Many people feel that if proper allowance is made for possible measurement error, no fixed accuracy ratio should be set.

9-7-2 Methods of Precision Measurement

Three Basic Measurements. Standards laboratories make three basic measurements on impedance standards which are, in the order of decreas-

ing accuracy:

1. Intercomparison of standards of approximately the same value
2. Scaling impedance by multiples of 10
3. Measurement of odd values

While the first is the most precise, in some ways it is the easiest, for usually *direct substitution* measurements are made in which first a standard and then an unknown are measured at the same terminals of the measuring instrument with no change in the setting of the coarse digits of the main adjustment. The measurement device need not contain accurate components as long as the difference measured is small, the short-term stability of the system is high, and the resolution and sensitivity are adequate. Scaling and odd value measurements are usually more difficult, depending on the type of impedance being measured and the accuracy required.

Precision Resistance Measurements [19, 55, 56]. One-to-one resistance comparisons are most often made on a Kelvin direct-reading ratio set, such as the one shown in Fig. 9-19. Both resistors being compared are measured on the same terminals with an auxiliary standard or "tare" on the other side of the bridge. Because the Thomas resistor has a 1-Ω value, the highest precision is required at this low value and, as a result, the lead and yoke adjustments (see Sec. 9-4) are particularly important if low-valued ratio arms are used for optimum overall sensitivity. Variations in the switch resistance in the adjustable ratio arm would be a major problem if conventional decade adjustments were used. Instead, these adjustments are of the shunt type, which places relatively large switched resistance in parallel with a low-valued fixed resistor so that switch resistance is much less important. This limits the adjustment range to a rather small deviation, and as a result, such ratio sets can only measure resistors of approximately the same value as the standard used.

Decade scaling of resistance calibrations can be performed by connecting 10 standards of one value in series or by a 10:1 ratio in a ratio set that is calibrated by such a series buildup. The error for such a procedure is the error in the 10:1 ratio multiplied by the number of factors of 10 required, and 2 ppm per decade is considered a reasonable error. Recently transfer boxes [57] have become available that make an extremely accurate 100:1 ratio possible by providing a way to connect 10 resistors first in parallel and then in series with small connection errors. If any one resistor differs from the nominal value by the fractional amount δ, the theoretical fractional error in ratio is approximately $9\delta^2/100$. This procedure allows extreme precision and scaling from 1 Ω to 10 kΩ with ½-ppm precision.

Odd values of resistors are measured by several devices; a universal ratio set [55], a precision decade with individually measured and adjusted

resistors, or a bridge incorporating an adjustable voltage divider that can be precisely calibrated by ratio methods [58].

Precision ac resistance measurements can take advantage of the accuracy and stability of transformer ratio arms. While such bridges are often used for resistance thermometry, they have also been used for comparing standard resistors [59]. Transformer ratio arms have been used for precision dc resistance measurements [60], with a flux detector to determine any dc ampere-turn unbalance. This device is particularly suited for the accurate measurement of low-resistance shunts.

Precision Capacitance Measurements [61, 62]. All three types of measurements can be made on capacitors with a single instrument, namely, a high-resolution transformer-ratio-arm bridge (see Sec. 9-5). Three-terminal substitution measurements on 1,000-pF standards to 0.01 ppm are relatively easy with the use of commercial bridges of this type. Ten-to-one ratios are estimated to be precise to approximately 1 ppm, and an odd-valued capacitor can be compared with the nearest standard capacitor with approximately the same precision. As a result of this precision, the main limitations become the stability of the standard capacitors themselves (particularly at high values of capacitance) and the sensitivity of the overall system at low values.

Precision Inductance Measurements. Inductance measurements are the least accurate of the three types of impedance parameters, but fortunately accuracy requirements for inductors are correspondingly low. One-to-one comparisons can be made on high-resolution Owen or Maxwell bridges to within 1 ppm at medium values, but scaling or odd-value measurements on such bridges are limited by the accuracy of the bridge arms and other bridge errors. Series connections of standard inductors are sometimes used to provide scaling, but residual lead inductance, shunt capacitance, and low Q values limit the accuracy.

Perhaps the most precise inductance measurements are those made on a precision capacitance bridge by measuring the change in the effective capacitance at the unknown terminals as an inductor is added in series or in parallel with a capacitor [63], but residual parameters and low Q values limit this method similarly.

Multiterminal Measurements. While three- and four-terminal measurements are commonly made on low-valued admittances and impedances respectively, many instruments contain even more terminals. Several wide-range instruments use five terminals so that three terminals can be used at one end of their range and four at the other. In some cases errors from series and shunt residuals are significant at the same time so that a combination five-terminal measurement is necessary. This occurs when extreme precision is required at a medium impedance level or at higher frequencies where series inductance and parallel capacitance

effects are both more critical. Small errors occur even in a five-terminal measurement because of interaction between series and shunt residuals. An eight-terminal measurement system has been proposed to make an almost perfect measurement [64].

At radio frequency it is difficult to make good multiterminal impedance-measuring circuits, and so two-terminal measurements are necessary. While such measurements are usually subject to error, they can be very accurate if a precision coaxial connector is used. Such a connector has a sharply defined reference impedance so that good repeatability is possible [65].

REFERENCES

1. Blackburn, J. F.: "Components Handbook," M.I.T. Radiation Laboratory Series, p. 65, McGraw-Hill Book Company, New York, 1949.
2. Willard, C. L.: "Resistance and Resistors," McGraw-Hill Book Company, New York, 1960.
3. Gibbings, D. L. H.: A Design for Resistors of Calculable AC/DC Resistance Ratio, *Proc. IEE (London), Pt. C*, vol. 110, no. 2, pp. 335–342, February, 1963.
4. Harris, F. K.: "Electrical Measurements," p. 214, John Wiley & Sons, Inc., New York, 1952.
5. *Ibid.*, p. 226.
6. Dummer, G. W. A., and H. M. Nordenberg: "Fixed and Variable Capacitors," McGraw-Hill Book Company, New York, 1960.
7. Golding, E. W.: "Electrical Measurements and Measuring Instruments," p. 126, Sir Isaac Pitman & Sons, Ltd., London, 1955.
8. Von Hippel, A. R.: "Dielectric Materials and Application," John Wiley & Sons, Inc., 1954.
9. Blackburn, J. F.: "Components Handbook," M.I.T. Radiation Laboratory Series, p. 115, McGraw-Hill Book Company, New York, 1949.
10. Golding, E. W.: "Electrical Measurements and Measuring Instruments," p. 170, Sir Isaac Pitman & Sons, Ltd., London, 1955.
11. *Ibid.*, p. 44.
12. Hersh, J. F.: Connection Errors in Inductance Measurement, *Gen. Radio Experimenter*, October, 1960.
13. Terman, F. E.: "Radio Engineers' Handbook," p. 948, McGraw-Hill Book Company, New York, 1943.
14. Alonzo, G. J., R. H. Blackwell, and H. V. Marantz: Direct Reading Fully Automatic Vector Impedance Meters, *Hewlett-Packard J.*, January, 1967.
15. Holtje, M. C., and H. P. Hall: A High-precision Impedance Comparator, *Gen. Radio Experimenter*, April, 1956.
16. Hague, B: "Alternating Current Bridge Methods," p. 2, Sir Isaac Pitman & Sons, Ltd., London, 1957.
17. Harris, F. K.: "Electrical Measurements," p. 282, John Wiley & Sons, Inc., 1952.
18. Warshawsky, I.: Multiple-bridge Circuits for the Measurement of Small Changes in Resistance, *Rev. Sci. Instr.*, vol. 26, p. 711, 1955.
19. Wenner, F.: Methods, Apparatus and Procedures for the Comparison of Precision Standard Resistors, *J. Res. NBS*, vol. 25, August, 1940.

20. Harris, F. K.: "Electrical Measurements," p. 698, John Wiley & Sons, Inc., 1952.
21. Hague, B.: "Alternating Current Bridge Methods," p. 285, Sir Isaac Pitman & Sons, Ltd., London, 1957.
22. Clark, H. A. M., and P. B. Wanderlyn: AC Bridges with Inductively Coupled Ratio Arms, *Proc. IEE (London)*, vol. 96, pp. 365–378, May, 1949.
23. Oatley, C. W., and J. G. Yates: Bridges with Coupled Inductive Ratio Arms as Precision Instruments for the Comparison of Laboratory Standards of Resistance or Capacitance, *Proc. IEE (London)*, vol. 101, pp. 91–100, March, 1954.
24. Lynch, A. C.: A Bridge Network for the Precise Measurement of Direct Capacitance, *Proc. IEE (London)*, *Pt. B*, vol. 104, p. 363, 1957.
25. Crawford, E. C.: Impedance Measurements and the In Situ Component Bridge, *Marconi Instrumentation*, vol. 9, no. 2, 1963.
26. Gibbings, D. L. H.: An Alternating-current Analogue of the Kelvin Double Bridge, *Proc. IEE (London)*, *Pt. C*, vol. 109, p. 307, 1962.
27. Hersh, J. F.: A Close Look at Connection Errors in Capacitance Measurements, *Gen. Radio Experimenter*, July, 1959.
28. Easton, I. G.: A Guard Circuit for the Capacitance Bridge, *Gen. Radio Experimenter*, August, 1952.
29. Hill, J. J., and A. P. Miller: An AC Double Bridge with Inductively Coupled Ratio Arms for Precision Platinum Resistance Thermometry, *Proc. IEE*, vol. 110, no. 2, p. 453, 1963.
30. Foord, T. R., R. C. Langlands, and A. J. Binnie; Transformer-ratio Bridge Network with Precise Lead Compensation, *Proc. IEE (London)*, vol. 110, no. 9, September, 1963.
31. Hall, H. P.: The Measurement of Electrolytic Capacitors, *Gen. Radio Experimenter*, June, 1966.
32. Hague, B.: "Alternating Current Bridge Methods," p. 600, Sir Isaac Pitman & Sons, Ltd., London, 1957.
33. *Ibid.*, p. 468.
34. Logan, M. A.: An AC Bridge for Semiconductor Resistivity Measurements Using a Four-point Probe, *Bell System Tech. J.*, vol. 40, no. 3, pp. 885–920, May, 1961.
35. Hague, B.: "Alternating Current Bridge Methods," p. 297, Sir Isaac Pitman & Sons, Ltd., London, 1957.
36. Yoshimoto, K.: A New Universal Bridge with Simplified, Semi-automatic Tuning, *Hewlett-Packard J.*, vol. 18, no. 1, August, 1966.
37. Fulks, R. G.: The Automatic Capacitance Bridge, *Gen. Radio Experimenter*, April, 1965.
38. Soderman, R. A.: New RF Bridge Features Small Size and Added Operating Convenience, *Gen. Radio Experimenter*, June, 1955.
39. Mennie, J. H.: A Wide-range VHF Impedance Meter, *The Boonton Radio Co. Notebook*, Summer, 1954.
40. Calvert, R.: U.S. Patent 2,589,535, Feb. 13, 1950.
41. Thurston, W. R.: A Direct-reading Impedance Measuring Instrument for the VHF Range, *Gen. Radio Experimenter*, May, 1950.
42. Tuttle, W. N.: Bridged-T and Parallel-T Null Circuits for Measurements at Radio Frequencies, *Proc. IRE*, vol. 23, no. 2, p. 88, February, 1935.
43. Sinclair, D. B.: The Twin-T—A New Type of Null Instrument for Measuring Impedance at Frequencies up to 30 Megacycles, *Proc. IRE*, vol. 28, no. 7, p. 310, July, 1940.
44. Huntley, L. E.: A Self-calibrating Instrument for Measuring Conductance at Radio Frequencies, *J. Res. NBS C*, vol. 69, April–June, 1965.

45. Sinclair, D. B.: Parallel-resonant Methods for Precise Measurements of High Impedance at Radio Frequencies and a Comparison with Ordinary Series-resonant Methods, *Proc. IRE*, vol. 26, no. 12, p. 1466, December, 1938.
46. Dalke, J. L., and R. C. Powell: Measuring Power Factor of Low-loss Dielectrics, *Electronics*, vol. 24, no. 8, p. 224, August, 1951.
47. Cook, L. O.: A Versatile Instrument—The Q Meter, *The Boonton Radio Co. Notebook*, no. 4, Winter, 1955.
48. Silsbee, F. B.: Establishment and Maintenance of the Electrical Units, *NBS Circ.* 475, June, 1949 (available from the U.S. Government Printing Office, Washington, D.C.).
49. Silsbee, F. B.: Extension and Dissemination of the Electrical and Magnetic Units by the National Bureau of Standards, *NBS Circ.* 531, July, 1952.
50. Thompson, A. M., and D. G. Lampard: A New Theorem in Electrostatics and Its Application to Calculable Standards of Capacitance, *Nature*, vol. 166, p. 888, 1956.
51. Cutkosky, R. D.: Evaluation of the NBS Unit of Resistance Based on a Computable Capacitor, *J. Res. NBS, A*, vol. 65, p. 142, 1961.
52. Thompson, A. M.: An Absolute Determination of Resistance Based on a Calculable Standard of Capacitance, *Metrologia*, vol. 4, no. 1, p. 107, January, 1968.
53. NASA NTC 200-3, MIL-C-45662A, 1962.
54. "Standards Laboratory Information Manual," Naval Inspector of Ordnance, Pomona, Calif.
55. Thomas, J. L.: Precision Resistors and Their Measurements, *NBS Circ.* 470.
56. Harris, F. K.: "Electrical Measurements," p. 204, John Wiley & Sons, Inc., New York, 1952.
57. Hamon, B. V.: A 1–100 Ω Build-up Resistor for the Calibration of Standard Resistors, *J. Sci. Instr.*, vol. 31, no. 12, rec. 54, pp. 450–453, 1964.
58. Julie, L.: Ratio Metrics—A New, Simplified Method of Measurement Calibration and Certification, *IEEE 1964 Conv. Record, Pt. 8*, Instrumentation, p. 225, 1964.
59. Hill, J. J.: Calibration of DC Resistance Standards and Voltage Ratio Boxes by an AC Method, *Proc. IEE (London), Pt. 1*, vol. 112, no. 1, p. 211, June, 1963.
60. MacMartin, M. P., and N. L. Kusters: A Direct-current Comparator Ratio Bridge for Four-terminal Resistance Measurements, *IEEE Trans. Instr.*, vol. IM-15, no. 4, pp. 212–219, December, 1966.
61. McGregor, M. C., et al.: New Apparatus at NBS for Absolute Capacitance Measurement, *IRE Trans. Instr.*, vols. 1–7, December, 1950.
62. Hersh, J. F.: Accuracy, Precision and Convenience for Capacitance Measurement, *Gen. Radio Experimenter*, August–September, 1962.
63. Hillhouse, D. L., and J. W. Kline: A Ratio Transformer Bridge for Standardization of Inductors and Capacitors, *IRE Trans. Instr.*, vols. 1–9, no. 2, September, 1960.
64. Cutkosky, R. D.: Four-terminal Pair Networks as Precision Standards, *IEEE Trans. Commun. and Electron.* no. 20, p. 19, January, 1964.
65. Huntley, L. E., and P. N. Jones, Lumped Parameter Impedance Measurements, *Proc. IEEE*, vol. 55, no. 6, pp. 900–911, June, 1967.

CHAPTER TEN

AUDIO FREQUENCY SIGNAL SOURCES

Donald E. Norgaard

Hewlett-Packard Company, Palo Alto, California

Audio-frequency signal generators serve as convenient sources of electrical signals that are useful in the test, maintenance, or operation of a wide variety of electrical apparatus. Although the term *audio frequency* refers to the frequency range to which the ear is responsive, many audio-frequency signal generators are designed to produce signals both higher and lower in frequency than the usually accepted audio range from 20 to 20,000 Hz. In general this extension in range is accomplished by techniques identical with those employed for the audio range. It should be noted that the discussed generators provide electrical signals only, not acoustical signals.

10-1 The Sinusoidal Audio-frequency Signal Source

Alternating-current circuit theory has been developed around the concept of voltages, currents, and impedances stated in terms of a sinusoidal signal, a signal which has its energy concentrated at a single point in

the frequency domain. As a consequence, the majority of audio-signal generators in use are designed to provide sinusoidal output signals as nearly as possible. Stated more explicitly, the sinusoidal audio-signal generator should provide an output signal that has the following characteristics:

1. Low harmonic content
2. Stable operating frequency
3. Stable output amplitude
4. Low spurious output, i.e., hum, noise, jitter, modulation, etc.

The questions of how low the harmonic and spurious outputs should be and how stable the frequency and amplitude should be depend, of course, upon the application for the signal generator. Many applications do not require extremes in all four of the above characteristics simultaneously; some do. The general-purpose sinusoidal audio-frequency generator, however, should be capable of satisfying all four requirements reasonably well so that its range of application can be as broad as possible.

The general-purpose generator should also have provision for adjustment of its operating frequency over its entire range of capability and some means for adjustment of its output level as well. Operating frequency usually is adjusted by means of an analog dial or scale, although some designs employ push buttons along with an analog vernier which permits interpolation between the smallest steps obtainable with the push-button arrangement. Output-level control is generally accomplished by means of a continuously variable attenuator or by means of a step attenuator in conjunction with a continuously variable control for interpolation between steps of output level. In almost every case, the voltage level at which the signal is generated is constant; the output level is then controlled by attenuation rather than by control of the level of oscillation alone within the signal generator.

The maximum output power level obtainable from most general-purpose audio-signal generators is usually fairly low, ranging from 1 to perhaps 100 mW. Whenever higher power level is required, an external power amplifier is used to obtain the desired level. Some special-purpose signal generators are built to provide much greater power output. The additional capability is produced by a self-contained power amplifier which is, however, electrically separate from the generator itself.

The audio-frequency signal generator, in its generalized form, accepts electrical energy from some source such as a battery or rectifier and converts a portion of this energy into its output signal. This signal is a cyclic repetition of electrical voltage or current of predetermined frequency, waveform, and magnitude. The usefulness of the sinusoidal waveform in analysis and test of a wide variety of apparatus warrants

consideration of the several methods by which such signals are produced and controlled.

Practical sinusoidal signal generators can be classified by the method of frequency control into the following types: (1) the *LC* oscillator, (2) the *RC* oscillator, (3) the beat-frequency oscillator, and (4) special methods.

It will be noted that the term *oscillator* is introduced as a means of describing the first three classes listed. The logic of this terminology stems from the use of regenerative feedback to neutralize system losses and thereby produce a condition of sustained oscillation within the frequency-controlling elements.

10-2 The *LC* Oscillator

Historically, the *LC* oscillator came first. The analogy between a parallel-connected inductor and capacitor and the simple pendulum was recognized by Tesla and Marconi and utilized by each to produce oscillation at rather high power level. The oscillating system of inductor and capacitor was excited by spark discharge that caused damped trains of oscillation following each discharge. Later, Poulsen used a steady electric arc to produce undamped (continuous-wave) oscillation in the *LC* circuit, also at surprisingly high power levels.

The vacuum tube, with its capability of power gain, was applied by Armstrong to supply energy lost within a continuously oscillating *LC* circuit by a regenerative association of the tube with the circuit. The low-power signal generator became a practical reality with this combination.

The impedance characteristics of a parallel connection of lossless reactive elements L and C uniquely define a point in the frequency domain where the antiresonant circuit so formed exhibits infinite impedance. Realizable reactive circuit elements have losses which alter the impedance significantly at frequencies near the antiresonant frequency. The losses in realizable reactances can be simulated by considering a theoretically lossless inductor-capacitor (*LC*) combination shunted by a resistor of such a value that power dissipated in the resistor at a given voltage level is equal to the total of that dissipated in imperfect reactive circuit elements operating at the same voltage. Use of this fiction facilitates analysis of the impedance characteristics of a physically realizable antiresonant circuit when it is used in conjunction with other circuit elements in a working system.

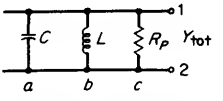


FIG 10-1 Generalized antiresonant circuit.

The characteristics of the elemental circuit diagram of Fig. 10-1 approximate reality closely enough to permit delineation of the properties

of antiresonance which enable the LC circuit to serve as a frequency-control mechanism in a signal generator. The following definitions are useful: L = inductance in henries, C = capacitance in farads, R_p = resistance in ohms, $\omega = 2\pi$ (frequency), and $j = \sqrt{-1}$.

$$Y_c = \frac{1}{R_p} \tag{10-2-1}$$

$$Y_b = -j\frac{1}{\omega L} \tag{10-2-2}$$

$$Y_a = j\omega C \tag{10-2-3}$$

$$Y_{\text{tot}} = Y_a + Y_b + Y_c = j\omega C - j\frac{1}{\omega L} + \frac{1}{R_p}$$

$$= \frac{\omega L + jR_p(\omega^2 LC - 1)}{\omega L R_p} \tag{10-2-4}$$

If we let $LC = 1/\omega_0^2$, then

$$Y_{\text{tot}} = \frac{\omega L + jR_p(\omega^2/\omega_0^2 - 1)}{\omega L R_p} \tag{10-2-5}$$

If Z is the impedance seen between points 1 and 2 at the frequency $\omega/2\pi$, then

$$Z = 1/Y_{\text{tot}} = \frac{R_p}{1 + j(R_p/\omega L)(\omega^2/\omega_0^2 - 1)} \tag{10-2-6}$$

Equation (10-2-6) expressed in polar form becomes

$$Z = |Z_\omega| \underline{/\phi_z} = |Z| \underline{/\tan^{-1} \frac{-R_p(\omega^2/\omega_0^2 - 1)}{\omega L}} \tag{10-2-7}$$

The normalized phase slope is

$$\frac{d\phi_z}{d(\omega/\omega_0)} = -\frac{2\omega^2 L R_p/\omega_0 - \omega_0 L R_p(\omega^2/\omega_0^2 - 1)}{\omega^2 L^2 + R_p^2(\omega^2/\omega_0^2 - 1)} \tag{10-2-8}$$

$$\frac{d\phi_z}{d(\omega/\omega_0)} = -\frac{2R_p}{\omega_0 L} \qquad \text{when } \omega/\omega_0 = 1 \tag{10-2-9}$$

From Eq. (10-2-7)

$$\phi_z = \frac{\pi}{4} \quad \text{rad}$$

at $\omega = \omega_1$ when ω_1 is determined as follows:

$$R_p(1 - \omega_1^2/\omega_0^2) = \omega_1 L \tag{10-2-10}$$

Solving for ω_1,

$$\omega_1 = \frac{-L\omega_0^2/R_p + \sqrt{L^2\omega_0^4/R_p^2 + 4\omega_0^2}}{2} \tag{10-2-11}$$

Similarly, ϕ_z is $-\pi/4$ rad at $\omega = \omega_2$ when

$$\omega_2 = \frac{L\omega_0^2/R_p + \sqrt{L^2\omega_0^4/R_p^2 + 4\omega_0^2}}{2} \tag{10-2-12}$$

It follows that the geometric mean of ω_1 and ω_2 is

$$\sqrt{\omega_1\omega_2} = \omega_0 \tag{10-2-13}$$

while the difference between ω_2 and ω_1 is

$$\omega_2 - \omega_1 = \omega_0^2 \frac{L}{R_p} \tag{10-2-14}$$

The ratio of the geometric mean to the difference $(\omega_2 - \omega_1)$ is

$$\frac{\sqrt{\omega_1\omega_2}}{\omega_2 - \omega_1} = \frac{R_p}{\omega_0 L} \tag{10-2-15}$$

Equation (10-2-15) may be considered the ratio of the effective parallel resistance R_p to the reactance of either of the reactive elements of Fig. 10-1 when $\omega = \omega_0$. Thus is defined $Q_0 = R_p/\omega_0 L$, where Q_0 is the quality factor Q of the antiresonant circuit of Fig. 10-1 at $\omega = \omega_0$.

In terms of Q_0, the normalized phase slope is

$$\frac{d\phi_z}{d(\omega/\omega_0)} = -2\frac{R_p}{\omega_0 L} = -2Q_0 \tag{10-2-16}$$

Thus, when $\omega = \omega_0$ and ω_1 and ω_2 are defined as in Eqs. (10-2-11) and (10-2-12),

$$\frac{\sqrt{\omega_1\omega_2}}{\omega_2 - \omega_1} = Q_0 = -\frac{d\phi_z}{2d(\omega/\omega_0)} \tag{10-2-17}$$

The relationship between Q_0 and normalized phase slope will be used in analyses of other types of oscillators.

The circuit arrangement of Fig. 10-1 can be made self-oscillating by supplying power dissipated in R_p at a chosen operating level with the output of an amplifier whose input is derived from the signal appearing between terminals 1 and 2. The frequency of oscillation will depend upon the values of L and C and upon the phase shift within the amplifier. Such an oscillator is illustrated in Fig. 10-2, where R_p' represents the effective parallel resistance including the effect of the output impedance

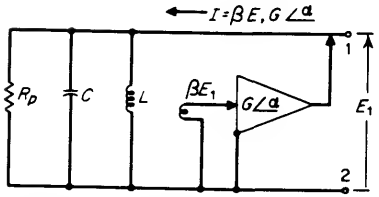


FIG 10-2 *LC* oscillator.

of the amplifier and a resistive load. The driving signal for the amplifier is a fraction of the signal appearing across the output of the amplifier, which has a mutual conductance of $G\underline{/\alpha}$.

If it is assumed that the input signal for the amplifier is βE_1, where β is a real number, then for steady-state conditions,

$$G\underline{/\alpha} \times \beta E_1 |Z| \underline{/\phi_z} = E_1 \tag{10-2-18}$$

where $|Z|\underline{/\phi_z}$ is the impedance seen between points 1 and 2 at the operating frequency. Equation (10-2-18) may be expressed as

$$G\underline{/\alpha}|Z|\underline{/\phi_z} = \frac{1}{\beta} \tag{10-2-19}$$

The voltage gain of the amplifier when driving its load $|Z|\underline{/\phi_z}$ is

$$K = G\underline{/\alpha}|Z|\underline{/\phi_z} = \frac{1}{\beta} \tag{10-2-20}$$

a real number. The phase angle α associated with G must be matched by an equal but opposite phase angle ϕ_z associated with Z. Such a phase angle attributed to G automatically causes the frequency of oscillation to adjust itself so that the product $G\underline{/\alpha}|Z|\underline{/\phi_z}$ is a real number.

Equations (10-2-6) and (10-2-7) indicate that $-\pi/2 < \phi_z < \pi/2$. Therefore, the allowable range of α lies between $\pm\pi/2$ rad if the system is to be capable of self-sustained oscillation.

Since the frequency of oscillation departs from the nominal frequency $f_o = \omega_0/2\pi = 1/2\pi\sqrt{LC}$ in order to accommodate the phase angle α, frequency stability (for a given LC product) is directly related to the stability of α and to the slope of the phase angle ϕ_z, which slope is a direct function of the circuit Q. The normalized phase slope numerically is greatest at $\omega/\omega_0 = 1$ where ϕ_z itself is zero. Instabilities in α will have the least effect upon the frequency of oscillation when the nominal value of α is zero.

Equation (10-2-19) calls for a precisely determined value of $G\underline{/\alpha}$ for

any stated values of $|Z|\underline{/\phi_z}$ and β under a steady-state condition of oscillation. A value of $|GZ|$ lower than stated in Eq. (10-2-19) precludes the possibility of sustained oscillation, while a greater value of $|GZ|$ would result in an ever ascending level of oscillation. It is inconceivable that the level of oscillation could continue to increase without limit in any real system; eventually a limit must be reached. Thus, to be useful, the oscillator system must exhibit an initial value of $|GZ|$ greater than stated in Eq. (10-2-19) so that oscillation can build up from the level of thermal noise when starting, and then $|GZ|$ must decrease to satisfy the equation at some predetermined level.

In most LC oscillator designs amplitude limiting occurs because G eventually becomes a decreasing function of the magnitude of either the input signal or the output signal of the amplifier. Energy storage in the reactive circuit elements L and C of Fig. 10-2 tends to keep the output signal E_1 nearly sinusoidal if $Q_0 \gg 1$, while the energy dissipated in R'_p is supplied during the cycle of oscillation from that stored in those elements. The energy lost by these elements must be replaced by the amplifier at least once for each cycle of operation if truly steady-state conditions are to be maintained. However, because of energy storage in L and C, this replacement energy need not be put into the oscillating system of L, C, and R'_p on an instantaneous supply-and-demand basis. Therefore, the amplifier can have a nonlinear G characteristic without serious effect on the nature of the output voltage E_1. The average value of G, however, must depend upon the average output voltage level in such a manner that Eq. (10-2-19) is satisfied on an average basis at some reasonable amplifier-output-voltage level.

The output current of the amplifier under steady-state conditions usually takes the form of a relatively narrow fixed-amplitude pulse repeated each cycle. Averaging of G is accomplished by variations of the area of the current pulse as energy requirements demand. This mode is known as *class C operation* and is a mode well suited for the LC oscillator. It is axiomatic that the voltage E_1 cannot be an absolutely pure sinusoid, since the current pulse which drives the antiresonant circuit is relatively rich in harmonic content. Although the antiresonant circuit acts as a harmonic filter, its attenuation of harmonic signals is finite.

A more consequential result of distortion generated within the amplifier is the dependence of amplifier phase shift upon the above mechanism of amplitude stabilization. Any condition that necessitates an adjustment of average G for the sake of amplitude stabilization usually results in a change of effective phase shift within the amplifier due to change of shape of the amplifier output current pulse or its time relationship to the input signal βE_1. The operating frequency then responds to this phase shift in order to satisfy Eq. (10-2-20).

Useful output power from the LC oscillator is available at the expense of frequency stability, since the effect of power extraction from the oscillating system is to lower the effective value of Q_0. In addition, any external loading that alters ω_0 owing to reactance coupled to the antiresonant circuit will perforce shift the operating frequency. Where frequency stability is a primary requirement, the LC oscillator must be lightly loaded. Good isolation between the oscillator and an external load may be obtained by means of an additional amplifier driven by a signal available from the oscillator itself.

The implementation of the LC oscillator shown in Fig. 10-2 is but one of many specific arrangements possible. The foregoing analysis can be applied to other configurations that utilize the impedance properties of an antiresonant LC circuit for frequency control. The essential principle in all regenerative LC oscillators is use of the power gain afforded by an amplifier to supply the power dissipated within the LC circuit in addition to that delivered to an associated load. Because an antiresonant LC circuit has the capability of discriminating against harmonic signals, distortion-free operation of the amplifier portion of the oscillator is not an important design consideration for most applications. As a consequence, regenerative LC oscillators serve as practical signal sources that provide useful output power with high efficiency in the conversion of input power to output signal power.

If it is assumed that the configuration shown in Fig. 10-1 is shunted by some device that exhibits the property of negative resistance, then a steady state of oscillation can be maintained when $(-R)^2 = (R_p)^2$, where $-R$ is the average value of the negative resistance postulated. Since the regenerative association of the amplifier with a similar LC circuit accomplishes the same result at some chosen signal level, the amplifier may be considered a source of negative resistance of exactly the same average magnitude as that of the positive resistance R'_p of Fig. 10-2.

Certain devices such as an electric arc, the magnetron, the tunnel diode, the dynatron, and the transitron exhibit negative resistance between two terminals and thus can be used to maintain oscillation in an LC circuit without the need for excitation such as the signal βE_1 of Fig. 10-2. These devices exhibit negative-resistance characteristics over only a portion of their possible voltage ranges; outside these ranges the characteristic resistance becomes positive. Thus, the average value of negative resistance so obtained is voltage dependent, with the result that the level of oscillation produced is a function of the equivalent positive resistance (such as R_p of Fig. 10-1) present in the circuit as well as the voltage-versus-current properties of the particular negative-resistance device. Under any circumstance, the absolute magnitude of negative resistance produced by such devices must be lower than the

equivalent positive resistance of the associated antiresonant LC circuit in order to allow oscillation to commence. Thereafter, the signal level builds up until the net average negative resistance offered by the negative-resistance device exactly balances the positive resistance of the remainder of the circuit.

Frequency Range. It is theoretically possible to choose LC products so that the operating frequency of LC oscillators can cover a very wide range. At low frequencies, however, practical problems associated with the reactive circuit elements, themselves, serve to limit the general usefulness of the LC oscillator. It has been shown that the frequency stability of the LC oscillator is a function of the effective Q of the antiresonant circuit as well as the stability of the LC product itself. Harmonic output is, in general, also a function of Q because of the pulselike nature of the current output of the amplifier that drives the antiresonant circuit. It is difficult to build inductors that have satisfactorily high values of Q at low frequencies, even when high-quality core materials are employed. In fact, when such materials are used in the effort to reduce both size and losses, saturation characteristics of the core materials become a major consideration in the overall design of a low-frequency LC oscillator.

It can be shown that when $\omega = \omega_0$,

$$\sqrt{L/C} = \frac{R_p}{Q_0} \tag{10-2-21}$$

Therefore, for given values of R_p and Q_0, the ratio L/C is predetermined. This ratio cannot be held constant in tunable oscillators unless both L and C are varied simultaneously. An approximation to simultaneous variation is achieved by providing a multiplicity of reactance elements which can be switched into use for coverage of a number of relatively narrow frequency ranges by variation of a single reactance element. At low frequencies the requirements for both L and C become ponderous. For example, at 16 Hz and with $Q_0 = 10$, then $R'_p = 1{,}000\ \Omega$ when $L = 1$ H and $C = 100\ \mu$F. If $L = 10$ H and $C = 10\ \mu$F, then $R'_p = 10{,}000\ \Omega$. The frequency and value of Q_0 are the same in both cases. An additional problem arises at low frequencies when it is desired to provide for continuous adjustment of the operating frequency, since conveniently adjustable inductors or capacitors of the general range of values indicated in this example are not readily available.

When LC oscillators are compared with other types later in the chapter, it will become clear why the kind discussed here find only limited use in variable-frequency audio generators. One reason is that a capacitance variation of 10:1 produces a frequency range of only $\sqrt{10}$:1 and it is difficult to achieve a greater range than that.

Radio-frequency signal generators of the LC type customarily combine an LC oscillator with an isolating amplifier whose output is fed to output terminals through an attenuator. Isolation of the oscillator from external loading achieved by this arrangement permits attainment of frequency stability approaching that afforded by the LC product itself. Spectral purity of the output signal can be enhanced by use of a tuned amplifier whose tuning is ganged with that of the oscillator, which attenuates harmonic energy present in the oscillator output. In addition, the amplifier may be designed to permit amplitude modulation by low-frequency signals. Radio-frequency signal generators of this type are widely used for test and alignment of receiving apparatus.

10-3 Resistance-Capacitance Signal Generators

Practical problems associated with the design of tunable LC circuits which operate in the audio-frequency range have led to the development of resistance-capacitance (RC) configurations for frequency control of signal generators operating in this range. A point of major difference between the properties of the antiresonant LC circuit and those of RC circuits used for frequency control is that use is made of the voltage transfer characteristics of RC networks rather than their impedance characteristics as was done in the case of the LC oscillator. Despite this difference in emphasis, there are many points of similarity between LC and RC oscillator systems that the following analyses serve to illustrate.

The RC circuit of Fig. 10-3 can be examined for its voltage transfer characteristics by considering the voltage ratio E_2/E_1 a function of the frequency of an applied voltage E_1. The upper portion of the network may be considered to have an impedance Z_1, while the lower portion has the impedance Z_2. Thus,

$$Z_1 = R - \frac{j}{\omega C_1} \tag{10-3-1}$$

$$Z_2 = \frac{-jR_2/\omega C_2}{R_2 - j/\omega C_2} \tag{10-3-2}$$

$$\begin{aligned} E_2/E_1 &= \frac{Z_2}{Z_1 + Z_2} \\ &= \frac{-jR_2/\omega C_2}{R_1R_2 - (1/\omega^2 C_1C_2) - j(R_1/\omega C_2 + R_2/\omega C_1 + R_2/\omega C_2)} \\ &= \frac{R_2/\omega C_2}{R_1/\omega C_2 + R_2/\omega C_1 + R_2/\omega C_2 + j(R_1R_2 - 1/\omega^2 C_1C_2)} \end{aligned} \tag{10-3-3}$$

When the quantity $R_1R_2 - 1/\omega^2C_1C_2 = 0$, then E_2 is in phase with E_1. For given values of R_1, R_2, C_1, and C_2 this occurs when

$$\omega^2 = \omega_0^2 = \frac{1}{R_1R_2C_1C_2} \tag{10-3-4}$$

which defines ω_0 in terms of circuit parameters.

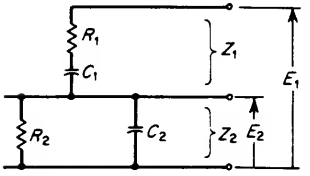


FIG 10-3 *RC* network.

Now, if $R_1 = kR_2$ and $C_2 = mC_1$, where k and m are real positive numbers, then Eq. (10-3-3) may be written

$$\frac{E_2}{E_1} = \frac{1}{k + m + 1 + j\sqrt{km}\,(\omega/\omega_0 - \omega_0/\omega)} \tag{10-3-5}$$

A plot of $|E_2/E_1|$ versus ω/ω_0 on semilogarithmic coordinates shows a maximum (or peak) at $\omega/\omega_0 = 1$ and geometric symmetry of $|E_2/E_1|$ about this peak.

Equation (10-3-5) may be expressed in generalized polar form

$$\frac{E_2}{E_1} = B\underline{/\phi_\omega} \tag{10-3-6}$$

where B is a positive real number and where

$$\phi_\omega = -\tan^{-1}\frac{\sqrt{km}\,(\omega/\omega_0 - \omega_0/\omega)}{k + m + 1} \tag{10-3-7}$$

A plot of ϕ_ω versus ω/ω_0 on semilogarithmic coordinates shows odd symmetry about $\omega/\omega_0 = 1$, at which value ϕ_ω is zero. Figure 10-4 is an illustrative plot of both E_2/E_1 and ϕ_ω versus ω/ω_0 for values of k and m both equal to unity.

The similarity of the voltage transfer characteristic exhibited by the *RC* network of Fig. 10-3 to the impedance characteristic of a parallel *LC* configuration of finite Q suggests an evaluation of the Q of the voltage transfer characteristic of the *RC* circuit.

For the *LC* configuration, the effective Q is stated as $(\omega_2\omega_1)^{1/2}/(\omega_2 - \omega_1)$, where ω_2 and ω_1 are, respectively, 2π times the frequency above and below resonance where the phase angle of the impedance has an absolute

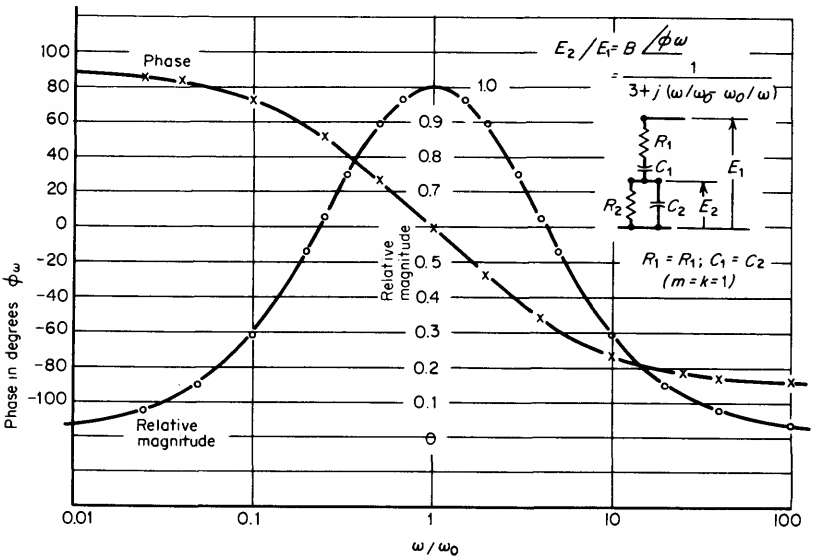


FIG 10-4 Phase- and normalized-amplitude characteristics of circuit in Fig. 10-3.

value of $\pi/4$ rad. The same reasoning applied to the RC circuit configuration of Fig. 10-3 involves determining ω_1 and ω_2 from Eq. (10-3-7) and substitution in the above expression for Q.

Above resonance, the phase angle of E_2/E_1 is $-\pi/4$ at ω_2 when

$$\sqrt{km}\left(\frac{\omega_2}{\omega_0} - \frac{\omega_0}{\omega_2}\right) = k + m + 1 \tag{10-3-8}$$

Solving for ω_2,

$$\omega_2 = \omega_0 \frac{(1/\sqrt{km})(k + m + 1) + \sqrt{[(1/km)(k + m + 1)]^2 + 4}}{2} \tag{10-3-9}$$

Similarly,

$$\omega_1 = \omega_0 \frac{-(1/\sqrt{km})(k + m + 1) + \sqrt{[(1/km)(k + m + 1)]^2 + 4}}{2} \tag{10-3-10}$$

The apparent Q of the transfer characteristic is

$$Q = \frac{\sqrt{\omega_2\omega_1}}{\omega_2 - \omega_1} = \frac{\sqrt{km}}{k + m + 1} \tag{10-3-11}$$

$Q = \frac{1}{3}$ when $k = m = 1$.

The normalized slope of the phase characteristic may be obtained by differentiating Eq. (10-3-7) with respect to ω/ω_0:

$$\frac{d\phi_\omega}{d(\omega/\omega_0)} = \frac{d \tan^{-1}[\sqrt{km}\,(\omega/\omega_0 - \omega_0/\omega)/(k+m+1)]}{d(\omega/\omega_0)} \tag{10-3-12}$$

$$\frac{d\phi_\omega}{d(\omega/\omega_0)} = \frac{-1}{1 + [km(\omega/\omega_0 - \omega_0/\omega)^2/(k+m+1)^2]} \frac{\sqrt{km}\,(1+\omega_0/\omega^2)}{k+m+1} \tag{10-3-13}$$

At $\omega/\omega_0 = 1$, in the region of interest,

$$\frac{d\phi_\omega}{d(\omega/\omega_0)} = \frac{-2\sqrt{km}}{k+m+1} \tag{10-3-14}$$

Thus, from Eq. (10-3-11),

$$\frac{d\phi_\omega}{d(\omega/\omega_0)} = -2Q \tag{10-3-15}$$

The similarity of Eq. (10-3-15) to Eq. (10-2-16), which applies to the impedance characteristic of the antiresonant LC circuit, is apparent.

10-4 The Wien Bridge Oscillator [1]

Figure 10-5 illustrates the RC circuit arrangement of Fig. 10-3 in combination with a resistive voltage divider comprising resistors R_3 and R_4. This circuit arrangement may be considered a bridge excited

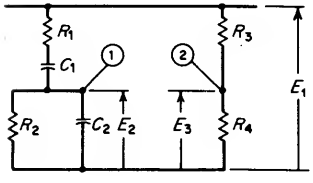


FIG 10-5 Wien bridge.

with the voltage E_1, with output available between points 1 and 2. From Eq. (10-3-5),

$$E_2 = \frac{E_1}{k+m+1+j\sqrt{km}\,(\omega/\omega_0 - \omega_0/\omega)} \tag{10-4-1}$$

$$E_3 = E_1 \frac{R_4}{R_3+R_4} = E_1\beta \tag{10-4-2}$$

where β is a real positive number ≤ 1.

The bridge output is

$$E_2 - E_3 = E_1\left[\frac{1}{k + m + 1 + j\sqrt{km}\,(\omega/\omega_0 - \omega_0/\omega)} - \beta\right] \tag{10-4-3}$$

If

$$\beta = \beta_0 - \Delta \tag{10-4-4}$$

where

$$\beta_0 = \frac{1}{k + m + 1} \tag{10-4-5}$$

is substituted in Eq. (10-4-3), then

$$E_2 - E_3 = E_1\left[\frac{1}{k + m + 1 + j\sqrt{km}\,(\omega/\omega_0 - \omega_0/\omega)} - \frac{1}{k + m + 1} + \Delta\right] \tag{10-4-6}$$

It follows that

$$\frac{E_2 - E_3}{\Delta} = E_1\left[\frac{(k + m + 1)^4 - km(1/\Delta - 1)(\omega/\omega_0 - \omega_0/\omega)^2 - j\frac{1}{\Delta}\sqrt{km}\,(k + m + 1)^2(\omega/\omega_0 - \omega_0/\omega)}{(k + m + 1)^4 + km(\omega/\omega_0 - \omega_0/\omega)^2}\right] \tag{10-4-7}$$

Equation (10-4-7) may be written in generalized polar form

$$\frac{E_2 - E_3}{\Delta} = G\underline{/\phi_B} \tag{10-4-8}$$

where G is a real number and where

$$\phi_B = \tan^{-1} - \frac{(1/\Delta)\sqrt{km}\,(k + m + 1)^2(\omega/\omega_0 - \omega_0/\omega)}{(k + m + 1)^4 - km(1/\Delta - 1)(\omega/\omega_0 - \omega_0/\omega)^2} \tag{10-4-9}$$

When $\omega/\omega_0 = 1$, the normalized phase slope $d\phi_B/d(\omega/\omega_0)$ may be written

$$\frac{d\phi_B}{d(\omega/\omega_0)} = \frac{2}{\Delta}\frac{\sqrt{km}}{(k + m + 1)^2} \tag{10-4-10}$$

Now, by substitution from Eq. (10-3-11),

$$\frac{d\phi_B}{d(\omega/\omega_0)} = -\frac{2}{\Delta}\frac{Q}{(k + m + 1)} \tag{10-4-11}$$

Thus, the effective Q of the bridge arrangement Q_B is

$$Q_B = \frac{Q}{\Delta(k + m + 1)} = \frac{\beta_0}{\Delta} Q \tag{10-4-12}$$

by equivalence with Eq. (10-3-15).

Equation (10-4-12) shows that the apparent Q of the RC network of Fig. 10-3 as evaluated in terms of the normalized phase slope at $\omega/\omega_0 = 1$ is altered by a factor β_0/Δ, when this network is considered a portion of a bridge arrangement such as in Fig. 10-5. This statement applies only when the evaluation is made in terms of the output of the bridge between points 1 and 2.

When the quantity Δ is made very small (the bridge is very nearly balanced), Q_B can become quite large. In this manner the point $\omega/\omega_0 = 1$ is sharply defined by the phase of the signal appearing between points 1 and 2 relative to that of the exciting signal E_1.

The Wien bridge arrangement of Fig. 10-5 can be made self-oscillating when an amplifier whose input signal is $E_2 - E_3$ is arranged to provide the bridge-exciting signal E_1 as its output. Figure 10-6 illustrates this implementation of the Wien bridge.

From Eq. (10-4-7), when $\omega/\omega_0 = 1$,

$$\frac{E_2 - E_3}{\Delta} = E_1 \tag{10-4-13}$$

Thus, for self-sustaining oscillation at $\omega/\omega_0 = 1$, the voltage gain of the amplifier must be

$$K = \frac{1}{\Delta} = \frac{E_1}{E_2 - E_3} \tag{10-4-14}$$

Since the gain K can be made very large, the quantity Δ can be correspondingly small.

Under conditions of steady-state self-oscillation, Eq. (10-4-14) must be satisfied exactly; too great a value of Δ precludes self-oscillation, while too small a value will cause the strength of the oscillation to increase to a level at which amplifier overload and accompanying distortion serve to make the average voltage gain of the amplifier conform exactly with that stated.

It is apparent that the use of strictly linear bridge elements in the oscillator arrangement of Fig. 10-6 would result in an impractical system; either the system would not oscillate at all (Δ too large), or it would oscillate at an ever increasing level (Δ too small) ultimately limited by the distortion characteristics of the amplifier.

W. R. Hewlett [2] solved this enigma by providing one of the arms of

the bridge with an element (R_4 of Fig. 10-6) whose resistance is a function of the voltage across it, while still making this element an almost perfectly linear one during the period of the intended oscillation. Hewlett used a tungsten-filament lamp which has a positive temperature coefficient of

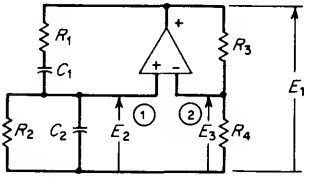


FIG 10-6 Wien bridge oscillator.

resistance. The self-heating of the filament due to flow of signal-frequency current serves to control the average resistance of R_4 so that β of Eq. (10-4-2) is dependent upon signal level. In this manner, Δ becomes dependent upon the operating level of the system, so that Eq. (10-4-14) is satisfied at a relatively fixed bridge signal level E_1. This operating level can be chosen to be well below the overload level of the amplifier.

Alternatively, a temperature-sensitive resistive element having a *negative* temperature coefficient of resistance can be used for the bridge element R_3. Thermistors are circuit elements of this class.

The practical significance of Hewlett's arrangement is to permit the amplifier voltage gain to be made very high, which forces the bridge to operate very near balance at a predetermined excitation level. At this level and with high gain, the resulting small unbalance provides a very high value of Q_B. Equation (10-4-14) is thus satisfied automatically, even though the value of amplifier gain is not precisely controlled.

Although relatively low values of K still permit oscillation to take place, the stability of the oscillating system is improved as the magnitude of K is increased. If amplifier gain should change, a readjustment of Δ automatically occurs because of a small change in operating level of the bridge. This change in operating level is small if K is large. Likewise, any phase shift in the amplifier is compensated for by an exactly equal but opposite phase shift in $E_2 - E_3$. The phase shift in $E_2 - E_3$ is brought about by a departure of the operating frequency from the nominal value of $\omega/\omega_0 = 1$ by just the correct amount to satisfy Eq. (10-4-14). A high value of phase slope (hence Q_B) in the bridge portion of Fig. 10-6 reduces the frequency departure necessary to compensate for amplifier phase shift.

Thus, small values of Δ resulting from large values of K serve to improve both amplitude stability and frequency stability of the system in terms of dependence on the stability of both gain and phase characteristics of the amplifier.

In general, amplifier phase shift is slightly different from zero. Therefore, the system usually does not operate at exactly ω_0, regardless of amplifier gain. Since the departure of operating frequency from the design value $\omega_0/2\pi$ is a function of amplifier gain, frequency stability is linked with gain stability of the amplifier at all operating points where amplifier phase shift is other than zero. Amplifier phase shift is generally greatest at frequencies beyond the limits of the *bandpass* characteristic, with the result that the relative frequency stability of the Wien bridge oscillator is best at operating frequencies well within the amplifier bandpass limits. Operation at frequencies outside the *cutoff* frequencies of the amplifier is possible, but carries with it the penalty of impaired frequency stability.

10-5 The Practical Wien Bridge Oscillator

The foregoing analysis purposely has evaded many of the practical aspects of a successful Wien bridge oscillator in order to concentrate on basic concepts of the system. The analysis has indicated that the operating frequency can be made to correspond closely with a selected design value and that the amplitude can be regulated to conform with a predetermined level despite variations in amplifier characteristics.

It is quite desirable that a general-purpose signal source be capable of operating over a wide range of frequencies and that provision be made for convenient selection of any chosen frequency of operation. In addition, it is desirable that the signal output be maintained at a selected level regardless of the frequency chosen. These two objectives are met by adjusting the parameters of the bridge as indicated by Eq. (10-3-4) for selection of operating frequency while preserving a constant ratio of both the resistance values R_1 and R_2 and the capacitance values C_1 and C_2 of the bridge throughout the entire range of adjustment of $R_1C_1R_2C_2$. Most designs call for simultaneous "tuning" of C_1 and C_2 or for simultaneous tuning of R_1 and R_2 while preserving their respective ratios. In the interest of both resettability and stability, the capacitors and resistors should be of good quality, i.e., mechanically and electrically stable.

A commonly used arrangement consists of a ganged configuration of air dielectric variable capacitors (C_1 and C_2) coupled to a dial mounted on the front panel. A frequency calibration is engraved on the dial. The range of capacitance variation is made somewhat greater than 10:1 so that a single set of calibration marks serves for several overlapping decade-related frequency ranges, which are selected in steps by switching fixed resistors (R_1 and R_2) simultaneously. The output of the oscillator is the bridge driving voltage E_1. This output may be taken directly through an attenuator to output terminals, or it may serve as a driving signal for a

power amplifier whose output is fed to the output system of the signal generator.

In some designs the input impedance of the amplifier might be low enough to disturb the bridge output signal, particularly at low frequencies where the RC products required cause the output impedance of the bridge to become large when relatively low capacitance values typical of air dielectric variable capacitors are involved. The output impedance of the bridge may be made appreciably lower than is practical with air dielectric capacitors by using a ganged pair of variable resistors for tuning. Frequency ranges are then selected by switching sets of fixed capacitors, essentially the reverse of the more commonly used arrangement described earlier.

Still other designs employ a series of push-button or dial-operated switches to change bridge parameters stepwise for frequency control. In these designs a ganged pair of variable resistors permits vernier interpolation between the smallest steps provided by the switching arrangement. The impedance level of the bridge can be controlled over a relatively wide range to suit amplifier input and output impedance characteristics.

Operation of the thermally controlled Wien bridge oscillator at frequencies lower than about 1 Hz is generally impractical because the thermal element is a source of odd-order harmonic distortion. If the thermal response speed is reduced significantly below that which causes more than about 1 percent distortion at 1 Hz, the amplitude control action becomes so sluggish that it is considered inadequate for most practical purposes.

The practical upper frequency limit of the Wien bridge oscillator is dominated by amplifier characteristics. It has been indicated that both high voltage gain and low phase shift at the operating frequency can be taken as measures of merit in such a system. The impedance $Z_1 + Z_2$ of Fig. 10-3 is

$$Z_1 + Z_2 = \frac{R_1(m + k + 1)(1 - j\sqrt{m/k})}{m + k} \tag{10-5-1}$$

at every value of ω where $\omega/\omega_0 = 1$. If $k = m = 1$, Eq. (10-5-1) becomes

$$Z_1 + Z_2 = 3\!/\!2\sqrt{2}\,R_1 \angle \frac{-\pi}{4} \tag{10-5-2}$$

If the minimum value of C_1 is 20 pF, then $X_{C_1} = R_1 = 800\ \Omega$ at 10^7 Hz when $k = m = 1$. Thus, when $R_1 = R_2 = 800\ \Omega$,

$$Z_1 + Z_2 = 1{,}200\sqrt{2} \angle \frac{-\pi}{4} \tag{10-5-3}$$

In addition to this reactive load, the amplifier must drive the remainder of the bridge ($R_3 + R_4$ of Fig. 10-6) and any external load as well. The oscillator operates at ω_0 only if the phase shift attributable to the amplifier (as it drives its load) is zero. Since it is unlikely that the amplifier will have zero phase shift at high frequencies, its voltage gain should be made as high as possible so that the bridge operates reasonably close to its condition of balance.

Amplifier gain and phase shift must be controlled outside the intended operating frequency range in the interest of satisfying the Nyquist stability criterion. Failure to do so can result in spurious oscillation which is not subject to control by the R_3, R_4 mechanism, and which is not closely related to the values of the frequency-determining bridge elements, R_1, R_2, C_1, and C_2. The frequency of the spurious oscillation is primarily related to amplifier phase shift, and only slightly influenced by bridge constants. Its amplitude, uncontrolled, drives the amplifier into saturation.

In general, at high operating frequencies, the performance of the Wien bridge oscillator is degraded by the compromises necessary to avoid spurious oscillation traceable to amplifier characteristics. The high-frequency limit of the Wien bridge oscillator is extended as new and better amplifier techniques and devices become available. The present practical frequency range from about 1 Hz to well over 1 MHz is great enough to satisfy most signal generator needs in the low-frequency portion of the spectrum.

10-6 *The Phase-shift Oscillator* [3,4]

The three-section low-pass ladder structure of resistance and capacitance shown in Fig. 10-7 exhibits a voltage transfer characteristic of such a nature that it may serve as a frequency-determining network in an oscil-

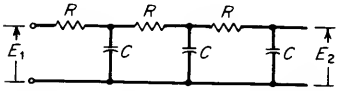


FIG 10-7 Three-section *RC* ladder.

lator. A ladder comprising three resistors of equal value R and three capacitors C has a voltage-transfer ratio of

$$\frac{E_2}{E_1} = \frac{1}{29/\underline{\pi}} \tag{10-6-1}$$

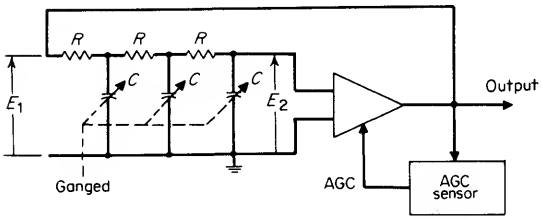


FIG 10-8 Ladder type of phase-shift oscillator.

at the frequency

$$f = \frac{\sqrt{6}}{2\pi RC} \tag{10-6-2}$$

This network will develop sustained oscillation at the design frequency when combined with a phase-inverting amplifier in the arrangement shown in Fig. 10-8, provided the amplifier has a voltage gain of exactly -29. (The minus sign indicates phase inversion within the amplifier.)

The frequency of oscillation may be varied by ganged variation of either the capacitors or the resistors of the ladder. It is required that the amplifier gain be closely regulated to exactly the value dictated by the attenuation of the ladder at the operating frequency. An automatic gain control (AGC) actuated by the output signal of the amplifier serves to establish a stabilized operating level below that at which amplifier overload causes excessive distortion in the output voltage E_1.

The low-pass nature of the RC ladder can itself serve as a harmonic filter by utilizing the signal E_2 as the input for a second low-distortion amplifier whose output serves as that of the signal generator.

The normalized phase slope of the network of Fig. 10-7 at the design frequency $f = \sqrt{6}/2\pi RC$ has a value of $-12\sqrt{6}/29$, which results in an apparent Q of 0.5 for the ladder type of phase-shift oscillator in Fig. 10-8. No multiplication of this apparent Q is inherent in the system, with the result that amplifier phase shift adds algebraically to that of the ladder network. Frequency stability, therefore, will be closely related to the phase stability of the amplifier at any fixed value of RC product used in the ladder.

10-7 The Ring Oscillator

Another implementation of the phase-shift oscillator concept is shown in Fig. 10-9. Three identical low-pass RC sections are isolated by phase-inverting amplifiers which serve also to overcome the attenuation attrib-

utable to each RC section. The combined effective phase shift of each RC section and its associated amplifier stage is 120°, which requires a phase shift of 60° in each RC section when three such sets of RC sections and amplifiers comprise a "ring." The frequency of oscillation of such an arrangement is $f = \sqrt{3}/2\pi RC$, if each amplifier produces exact phase inversion and a voltage gain of 2. As is true with any steady-state oscillating system, net system loss must be matched exactly by system gain. The voltage gain of each amplifier stage is controlled automatically so that the level of oscillation is stabilized at a value where amplifier distortion is acceptably low.

Frequency of oscillation is controlled by ganged variation of either the capacitance C or the resistance R of each phase-shifting section, as indicated in Fig. 10-9. Since the output signals of successive amplifiers of the oscillating ring are 120° apart in phase, the ring oscillator can provide three-phase output if two additional output amplifiers are excited by signals which appear at the outputs of the remaining two amplifiers of the ring.

The nominal voltage gain required of each amplifier is such that the ring type of oscillator is well adapted to operation at high frequencies where amplifiers of the required gain and phase characteristic are relatively simple.

As in the ladder type of phase-shift oscillator, phase shifts within the amplifiers cause the operating frequency to depart from the idealized value stated. The normalized phase slope of the frequency-determining system of the ring oscillator at its nominal operating frequency has a value of $-3\sqrt{3}/4$. The apparent Q of the ring oscillator in Fig. 10-9 is 0.65, 30 percent greater than that of the ladder type. No Q multiplication is inherent in the system.

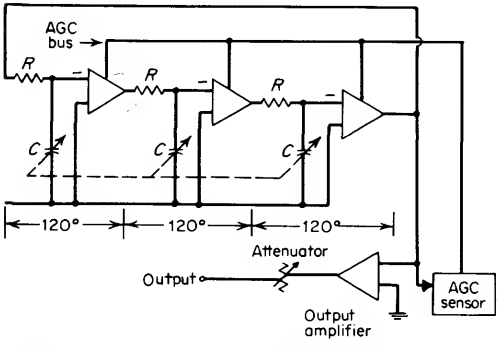


FIG 10-9 Ring type of phase-shift oscillator.

It is practical to provide a range of frequency adjustment over a ratio of greater than 10:1 for a given set of resistors used in conjunction with a three-gang variable air dielectric capacitor. Additional decade-related ranges may be provided by simultaneous switching of resistor sets in steps of 10:1 so that a single dial calibration serves for each of the decade-frequency ranges obtained in this manner.

Operation of the ring oscillator at low frequencies is beset with the difficulty of achieving low distortion concurrent with satisfactorily rapid automatic-gain-control action. Because of this practical limitation, ring oscillators are seldom designed for operation at frequencies lower than 10 kHz.

10-8 The Beat-frequency Oscillator

Two high-frequency oscillators whose difference frequency can be made to cover a desired low-frequency range serve as the basis of the beat-frequency oscillator. In its elemental form, the beat-frequency oscillator comprises two high-frequency oscillators, a mixer, a low-pass filter, and an output system. This is illustrated in block-diagram form in Fig. 10-10.

The mixer in Fig. 10-10 acts to perform a process of signal multiplication. In this case the signal output of oscillator 1 is in effect multiplied with that of oscillator 2. The following expressions describe the mixing process:

$$\text{Signal 1} = E_1 \sin \omega_1 t \tag{10-8-1}$$

$$\text{Signal 2} = E_2 \sin \omega_2 t \tag{10-8-2}$$

$$\text{Mixer output} = E_1 E_2 \sin \omega_1 t \sin \omega_2 t = \frac{E_1 E_2}{2} [\cos (\omega_1 - \omega_2)t - \cos (\omega_1 + \omega_2)t] \tag{10-8-3}$$

In practice, the magnitude of one of the two signals to be mixed together is made very much greater than that of the other with the result that the product may be shown to take the form

$$\text{Mixer output} = k \frac{E_1 E_2}{E_1} [\cos (\omega_1 - \omega_2)t - \cos (\omega_1 + \omega_2)t] + \text{dc component} + \text{high-frequency spurious components} \tag{10-8-4}$$

where k is a constant of proportionality, and E_1 is the larger of the two signals. The principal spurious components are those of frequency $\omega_1/2\pi$ and harmonics thereof, $\omega_2/2\pi$ and its harmonics, and a dc component.

The low-pass filter in Fig. 10-10 attenuates undesired high-frequency mixer products but passes the difference frequency $(\omega_1 - \omega_2)/2\pi$ and the

dc component to the output system. The output amplifier may be designed to reject the dc component, or alternatively, a balanced mixer may be used for suppression of the dc component. The practical mixer (whether balanced or not) serves to reduce the dependence of the output magnitude upon the strength of the larger of the two signals fed to the mixer. Thus, the low-frequency beat signal $kE_2 \cos(\omega_1 - \omega_2)t$ of Eq. (10-8-4) has constant magnitude when the output signal of oscillator 2 is maintained at a fixed level, even if the level of signal 1 is not constant.

Because the frequency of the desired output signal is the difference between those of the two high-frequency oscillators, the relative frequency stability of the beat signal is, in general, poorer than that of either of the two high-frequency sources. If, for instance, the desired beat frequency is 1 percent of the frequency of one of the high-frequency oscillators, a shift of 0.1 percent in the frequency of either oscillator results in a 10 percent shift of their difference frequency, a magnification of relative instability by a factor of 100. Very careful design of the high-frequency sources used in a beat-frequency system is necessary in order to stabilize the frequency of the beat signal to an acceptable degree.

The two high-frequency oscillators of the system must be nearly 100 percent free of mutual interaction in order to prevent "pulling" of one oscillator by the other. (Pulling is a tendency of the oscillators to synchronize with one another.) Such tendency causes phase modulation of the oscillators at their difference frequency and is a source of harmonic distortion in the resulting beat signal even though the two oscillators do not actually synchronize. Pulling effects between the two oscillators of a beat-frequency signal generator become more pronounced as their difference frequency decreases, with the result that low-frequency output signals of high-spectral purity and acceptable frequency stability are not generally practicable with the heuristically simple arrangement in Fig. 10-10.

It is necessary that $(f_1 - f_2) < f_2/2$, where $f_2 < f_1$. Consequently, several decades of output frequency range inherently are available in an

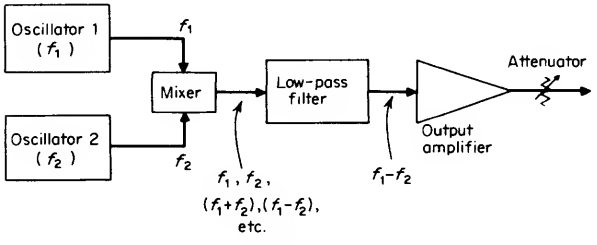


FIG 10-10 Beat-frequency oscillator.

unbroken continuum if desired. Such wide coverage (in frequency ratio) is available, however, at the expense of degraded frequency accuracy and also possible harmonic distortion in the low-frequency portion of the tuning range because of pulling.

Spurious signals incidental to generation of the desired output signal are suppressed only to the degree that the design of the low-pass filter permits. Many of the spurious signals are of significantly greater magnitude than the desired signal. Since these spurious signals are all higher in frequency than the desired signal, the lowest cutoff frequency of the filter consistent with uniform response over the desired range is indicated.

10-9 The Polyphase Beat-frequency Oscillator

The mixer output expressed in Eq. (10-8-4) was the result of mixing two signals having frequencies $\omega_1/2\pi$ and $\omega_2/2\pi$ respectively. If signals of the same two frequencies are mixed in a similar mixer, the difference frequency would be the same as that expressed in Eq. (10-8-4), but the phase of the output from the second mixer relative to that from the first will be determined by phase shifts introduced in the high-frequency signal paths feeding the mixers. Here is an analysis:

$$\begin{aligned}\text{Low-frequency output of mixer 1} &= kE_2 \cos (\omega_1 - \omega_2)t \\ &= kE_2 \cos (\omega_2 - \omega_1)t \end{aligned} \tag{10-9-1}$$

when the two signals applied were $E_1 \sin \omega_1 t$ and $E_2 \sin \omega_2 t$ respectively. In this case, all signals may be considered at "reference" phase.

An identical mixer supplied with signals $E_1 \sin \omega_1 t$ (reference phase) and $E_2 \sin (\omega_2 t + \alpha)$ (displaced α rad from its counterpart applied to mixer 1) will produce

$$\begin{aligned}\text{Low-frequency output from mixer 2} &= kE_2 \cos [(\omega_1 - \omega_2)t - \alpha] \\ &= kE_2 \cos [(\omega_2 - \omega_1)t + \alpha] \end{aligned} \tag{10-9-2}$$

The two low-frequency components expressed above differ only by the phase angle α introduced as a phase difference in one of the high-frequency mixer input signals.

The preservation of relative phase information throughout mixing processes is the basis for the polyphase beat-frequency oscillator [5]. Any desired number of separate output signals of the same frequency and independent phase relationship can be provided by expansion of the number of channels. The phase relationships established between these

channels are maintained throughout the entire range of output frequency when the high-frequency phase adjustments remain fixed.

The block diagram in Fig. 10-11 illustrates a two-phase beat-frequency signal generator that overcomes the principal disadvantages cited for the conventional beat-frequency oscillator conforming to the arrangement of Fig. 10-10. In this signal generator a number of overlapping output frequency ranges, each covering more than a decade, serve to permit operation at extremely low frequency without pulling and to provide adequately stable output frequency. A reference-phase signal is available from one of the two output channels provided. Signals of the same

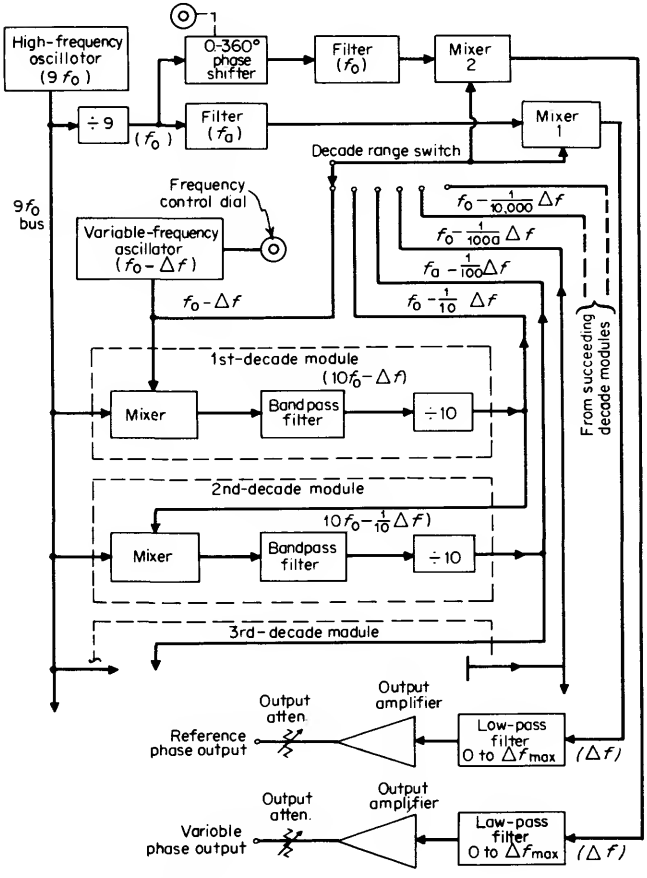


FIG 10-11 A polyphase beat-frequency oscillator.

frequency and of continuously adjustable phase are available from the other output channel.

Overlapping frequency ranges in decade steps are related to a primary range Δf, which is the highest of the group of ranges. This range is provided by mixing a high-purity amplitude-controlled signal of fixed frequency f_0 with the output signal of a buffered variable-frequency LC oscillator. The tuning range of the oscillator is about 10 percent of f_0 and this provides a beat-frequency ratio of about 12:1. The variable-frequency oscillator, therefore, is not required to approach the fixed frequency more closely than about 1 percent of f_0, which minimum separation is great enough to reduce pulling effects adequately.

The fixed-frequency signal at f_0 is split into two channels. One part is filtered and passed directly to a mixer supplied simultaneously with the output of the variable-frequency oscillator, while the other is supplied to a second mixer after passing through a phase-shifting goniometer and filter. The second mixer is also driven by the same variable-frequency signal as drives the first mixer. The goniometer is continuously variable over an unlimited range in either direction, lead or lag.

Succeeding beat-frequency ranges are produced by a group of mixers, filters, and frequency dividers in a series of iterative steps that effectively serve to decrease the beat-frequency range by successive decimal factors without, however, decreasing the original tuning range of the variable-frequency oscillator. Relative frequency stability then remains the same on all ranges. The method by which the secondary frequency ranges are achieved deserves further explanation, attempted below.

The fixed-frequency signal f_0 is obtained by frequency division from a stable high-frequency oscillator that operates at a fixed frequency $9f_0$. The signal at this frequency is used in the system to develop the remaining ranges as follows:

1. The signal at $9f_0$ is combined with the output of the variable-frequency oscillator in a mixer, shown in Fig. 10-11 as a portion of the first decade module.

2. A bandpass filter selects the sum term of the mixer output, $9f_0 + f_0 - \Delta f$, and supplies this signal, $10f_0 - \Delta f$, to a divide-by-10 frequency divider.

3. The output of the frequency divider, a frequency of $(f_0 - \Delta f/10)$, is brought to a range-selector switch and to the second decade module, where it is mixed with the $9f_0$ signal for a repeat of the process of mixing, filtering, and division by 10 as indicated.

4. Successive steps provide signals of $(f_0 - \Delta f/100)$, $(f_0 - \Delta f/1{,}000)$, and so forth, which are selected for mixing with f_0 in the two-channel mixers, as described earlier.

The process shown in Fig. 10-11 results in a series of exact decade steps

in output frequency so that a single calibration of the frequency control dial for the variable oscillator serves for all frequency ranges. Any number of decade ranges can be used to extend signal generation capabilities downward from the primary beat-frequency range.

Although signals of extremely low frequency can be generated by the above process, stabilized signal outputs from both reference-phase and variable-phase channels are available within a few microseconds after a change-of-range switch setting, or slewing, of either the frequency control or phase-shift control. Use of the variable-phase signal provides an operating convenience at very low frequencies where one cycle might have a duration of several minutes. The time axis of the signal can be positioned quite readily by use of the phase-shift control without the necessity of waiting. Throughout the entire range of frequency coverage the availability of two signals of the same frequency at any chosen phase relationship simplifies many measurements in which phase is a parameter of interest.

10-10 *Sine-wave Synthesis* [6]

A series of straight-line segments can be made to approximate an arbitrarily chosen waveform to as close a degree as desired if no limit is placed on the number of segments used. An approximation of a true sinusoid by as few as 26 linear segments of specifically selected lengths permits synthesis of the sinusoid within an rms error of less than 0.25 percent. Symmetry of the sinusoid about its coordinate axes requires use of only seven different lengths to perform straight-line synthesis within the error stated.

A sine wave may be synthesized from a symmetrical triangular wave to within a 0.25 percent rms error (distortion) by means of a series of 12 biased diodes which serve to switch 6 resistors in and out of an attenuating circuit at selected voltage levels, as shown in Fig. 10-12.

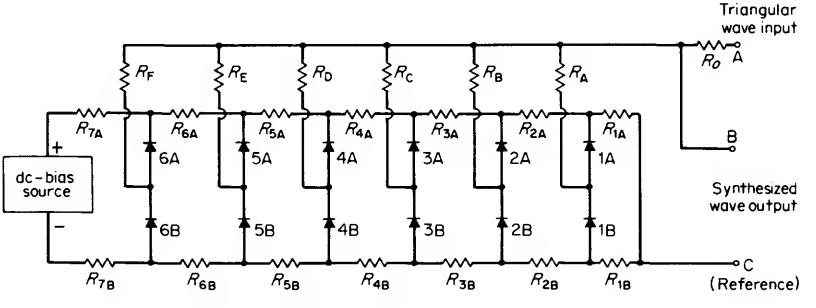


FIG 10-12 Triangular- to sine-wave synthesis.

The seven resistors R_{1A} through R_{7A} serve to bias the diodes $1A$ through $6A$ with respect to reference point C in a progression representing the approximate positive voltage levels of the required sinusoid where the successive linear segments are to be joined. Similarly, the seven resistors R_{1B} through R_{7B} serve to bias diodes $1B$ through $6B$ to the approximate negative voltage levels of the sinusoid where the successive segments join. A symmetrical triangular wave is applied between points A and C with the result that points A and B are at the same potential until current flows through either diode $1A$ or diode $1B$. For example, as the voltage of point A progresses in the positive direction from zero, the voltage level at which diode $1A$ begins to conduct marks the end of the first half segment of the approximation and the beginning of the second segment, since point B is now shunted to point C by the series combination of R_A, diode $1A$, and resistor R_{1A}, which forms an attenuator involving R_o. The same process is repeated progressively when diodes $2A$, $3A$, and so forth, conduct as the potential of point B becomes more positive. Following the positive crest of the triangular wave, diodes $6A$, $5A$, and so forth, progressively cease conduction as the voltage of point B becomes less positive. The negative portion of the sinusoid is synthesized in the same manner when diodes $1B$, $2B$, and so forth, progressively become conducting and serve to switch resistors R_A through R_F in and out of the system for the remainder of the wave.

It should be noted that the synthesis is optimized for minimum distortion of the synthesized signal for a specific set of fixed bias potentials and a related fixed voltage level of the applied triangular signal waveform. The synthesis is performed at as high a voltage level as feasible in order to minimize the relative effect of temperature variation on the characteristics of the diodes used in the process. The preordained voltage levels required by the synthesis demand that the triangular driving signal be closely maintained in symmetry, linearity, and magnitude at all operating frequencies.

A triangular wave that is suitable as a driving signal in the foregoing process is generated in the manner indicated in Fig. 10-13, in which a bistable (flip-flop) source that is capable of producing equal positive and negative potentials relative to a reference potential feeds its output to an adjustable potentiometer. Depending on the setting of this potentiometer, a certain fraction E_1 of the flip-flop output is supplied to a Miller integrator, which comprises resistor R_i, capacitor C_i, and a dc-coupled high-gain phase-inverting amplifier A, as shown in Fig. 10-13.

In the Miller integrator the voltage gain of the amplifier is made very high so that the potential variation at its input terminals is negligibly small compared with E_1. Under this circumstance a current $I_1 = E_1/R_i$ flows through R_i into point 1. If the input current to the amplifier is

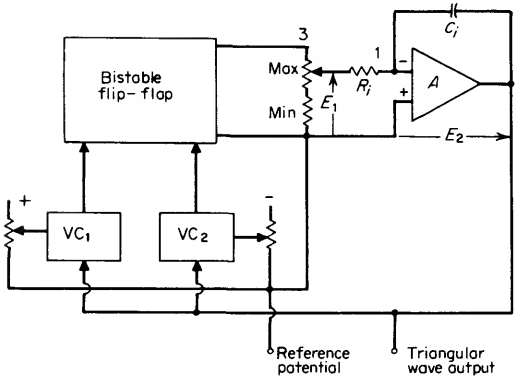


FIG 10-13 *Triangular-wave generator.*

negligibly small compared with I_1, essentially all of I_1 flows into (or out of) capacitor C_i, which causes the potential E_2 at the output terminal of the amplifier to change linearly with time at a fixed slope $dE_2/dt = I_1/C_i$.

The uniformly changing voltage E_2 is fed to two voltage comparators VC_1 and VC_2, which serve to trigger the bistable circuit to an opposite state and to reverse E_1 when the magnitude of E_2 reaches preset bias potentials of equal magnitude but opposite polarities with respect to the reference potential. Coincident with the polarity reversal experienced in E_1, the current I_1 reverses and thereby causes the potential E_2 to change linearly with a slope $dE_2/dt = -I_1/C_i$, which is equal but opposite to that which it has prior to the triggering event postulated. The potential E_2 maintains this opposite slope until the second voltage comparator is activated, at which time the bistable circuit is then triggered to its original state. The result is that E_2 becomes a wave of triangular shape having equal positive and negative slopes and also equal positive and negative voltage excursions relative to the reference potential.

Since the slope of the triangular wave is directly proportional to I_1, the repetition rate of the sequence just described is proportional to the magnitude of E_1, which permits the potentiometer to act as a frequency control for the system. Several tuning ranges can be provided by switching integrating capacitors of different values into and out of use as required.

The described mode of operation is often called *relaxation oscillation.* This mode is typified by one or more short bursts of activity per operating cycle, interspersed with periods of internal readjustment (usually charging or discharging of an RC circuit) in preparation for the next activity burst.

The triangular-wave generator combined with the triangular- to sine-

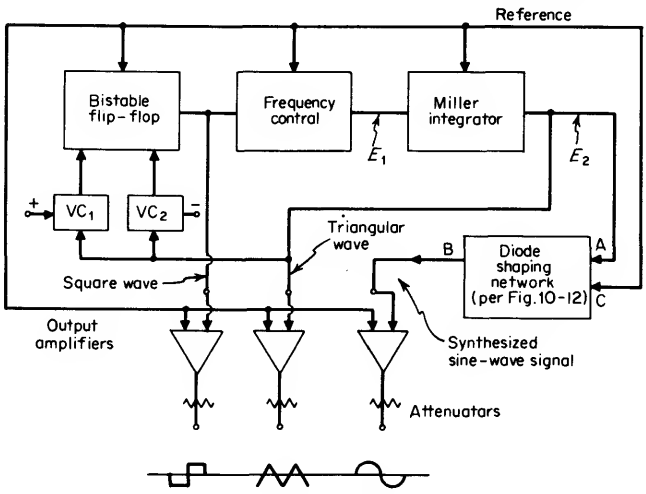


FIG 10-14 Function generator with synthesized sine-wave output.

wave synthesizer is illustrated in Fig. 10-14. Such a combination can serve as a multipurpose signal generator from which square-wave signals, sine-wave signals, and triangular-wave signals can be obtained either simultaneously through separate output systems (as shown) or selectively through a single output system. In such a configuration the triangular wave may be considered the integral of the square wave, while integration of the triangular wave an even number of times gives a sine wave as an approximation. Such a signal generator is commonly called a *function generator*.

Leakage resistance associated with the integrating capacitor C_i in Fig. 10-13 serves to place a limitation on the lowest frequency of operation practicable, although the use of high-quality capacitors permits successful operation at extremely low frequencies. The effect of shunt resistance across the capacitor is a gradual decrease in the magnitude of the slope of the triangular wave following each peak. The triangular wave thus loses symmetry and thereby degrades the quality of the synthesized sine wave as well. In addition, the period of the signal generated is increased, since triggering of the bistable circuit is delayed as a result of the decreased rate at which the triangular-wave signal approaches the preset triggering level in the voltage comparators.

Stray capacitance associated with the diodes and resistors of the synthesizing network in Fig. 10-12 serves to disturb the shaping process at high operating frequencies, with the result that additional distortion is introduced into the synthesized signal. In addition, maintenance of a

symmetrical triangular driving signal for the shaping network becomes increasingly difficult at high frequencies. The practical upper-frequency limit for synthesis depends upon choice of voltage and current levels used in the shaping network, the temperature stabilization achievable for the diodes, and control of stray reactance throughout the generating system.

A consequence of the synthesis method is that output amplitude is necessarily independent of operating frequency. In addition, rapid slewing of frequency (including range switching) cannot introduce a transient having a duration of more than a few microseconds following such intentional discontinuities.

The mechanism of frequency control employed in this type of signal generator permits the operating frequency to be varied by external electrical means. The frequency may be made a linear function of an external control voltage which serves to set the positive and negative excursions of the square-wave signal applied to the integrator as a substitute for the setting of the frequency control potentiometer shown in Fig. 10-13. The frequency can be "swept" by such a control voltage, which provides the capability of frequency variation over a range as great as several decades at a rate determined by the slope of the applied frequency control voltage. A repetitive control signal such as a sawtooth waveform permits successive sweeps of output frequency over a selected range. Such a signal at constant amplitude is useful in making rapid frequency-response measurements on many types of equipment.

CITED REFERENCES

1. Bauer, B.: Design Notes on the Resistance Capacity Oscillator Circuit, *Hewlett-Packard J.*, vol. 1, no. 3, November, 1949, and vol. 1, no. 4, December, 1949.
2. Terman, F. E., R. R. Russ, W. R. Hewlett, and F. C. Cahill: Some Applications of Negative Feedback with Particular Reference to Laboratory Equipment, *Proc. IRE*, vol. 27, pp. 649–655, October, 1939.
3. Nichols, H. W.: United States Patent 1,442, 781, Jan. 16, 1923 (filed July 7, 1921).
4. Ginzton, E. L., and L. M. Hollingsworth: Phase-shift Oscillators, *Proc. IRE*, vol. 29, pp. 43–49, February, 1941.
5. Crawford, R.: A Low Frequency Oscillator with Variable-phase Outputs for Gain-phase Evaluations, *Hewlett-Packard J.*, vol. 16, no. 11, July, 1965.
6. Brunner, R. H.: A New Generator of Frequencies Down to 0.01 CPS, *Hewlett-Packard J.*, vol. 2, no. 10, June, 1951.

CHAPTER ELEVEN

OSCILLOSCOPES

Charles H. House

Hewlett-Packard Company
Colorado Springs, Colorado

In the history of electrical and electronic measurement, no instrument has had greater impact than the *oscilloscope*. The ability to see phenomena happen has been virtually synonymous with the ability to measure their magnitude, and the development of all instrumentation has depended upon the ability of the measurement device to capture and display data for the operator's observation. Galvanometers and other mechanical apparatus capable of indicating dynamic phenomena existed long before the oscilloscope, but they were machines with slow response. The chief advantage of the early oscilloscope was its ability to display high-speed phenomena for easy observation, an asset which grew to ever greater importance as applications for electronics developed in the twentieth century. Today, the oscilloscope is the most versatile general-purpose electronic measuring instrument available for scientific investigation.

In 1879, William Crookes demonstrated the ability to deflect cathode rays in a vacuum tube with a magnet. Cathode rays had earlier been shown capable of causing phosphorescence on the glass walls of a vacuum tube, but control of the area of phosphorescence was only possible by using shaping masks (solid structures in the tube incapable of producing dynamic

deflection). The combination of focusing elements to produce a narrow electron beam (or cathode ray) aimed at a fluorescent target with dynamic electromagnetic-beam deflection became known as a *Crookes tube.* The Crookes tube, later to be commonly known as a *cathode-ray tube* (CRT) or more precisely as an *electron-beam tube,* offered much promise for displaying high-speed variations that could not be demonstrated on mechanical apparatus. By 1897, Karl F. Braun had constructed a "variable current apparatus" by using the Crookes tube, the first forerunner of the modern oscilloscope.

An oscilloscope is an instrument capable of presenting a luminous *xy* graph of any two related electrical parameters. One set of electrical signals is applied to the horizontal deflection (x-axis) system, and the other set of signals to the vertical deflection (y-axis) system, whereupon the cathode-ray beam traces out the *xy* coordinate graph. The intensity of the beam can be controlled, or modulated, by an electrical signal (sometimes called the *video signal*) applied to the z-axis system. Such a definition includes not only oscilloscopes and other related electronic instrumentation displays, but standard home television sets as well. In recent years, however, the definition of an oscilloscope has been altered in common usage to mean primarily an electronic instrument used for display of electrical signals on the vertical axis, compared with time on the horizontal axis. Thus the modern laboratory oscilloscope is a *time-domain* measurement and display instrument for analysis of electrical signals.

This chapter discusses several basic types of laboratory oscilloscopes and their component blocks, along with accessories that extend their usefulness in measuring mechanical, chemical, and medical, as well as electrical phenomena. The chapter also reviews representative oscilloscopelike instrumentation for many special-purpose measurements, not necessarily in the time domain.

11-1 The Oscilloscope Function

The function of the oscilloscope (scope) may be defined as *capturing, displaying,* and *analyzing* a time-domain waveform. All instrument developments since Braun's first oscilloscope are refinements in one or more of these three areas [1, 2, 3].

Traditionally, the oscilloscope has been used to capture high-speed transient phenomena and display them for the engineer or technician to analyze and interpret. Even in this simple operation, the oscilloscope (by its control settings) is performing a limited analysis function in conjunction with the operator. The great majority of scopes in use today are such general-purpose instruments; they capture a waveform and faith-

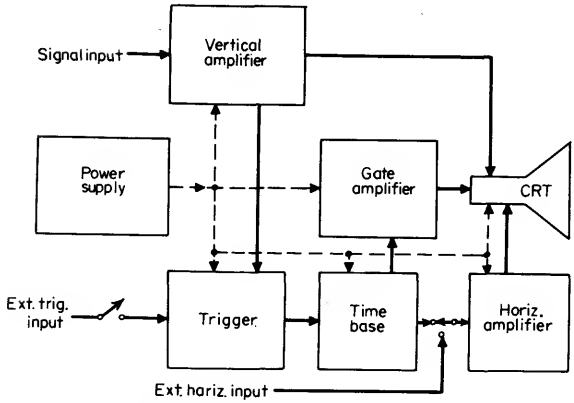


FIG 11-1 Oscilloscope block diagram.

fully portray that waveform on the CRT for the viewer to observe. Some oscilloscopes today are capable of analyzing data and presenting digital answers for their user; others, in conjunction with computers, are capable of wide-ranging data accumulation and analysis.

The modern time-domain oscilloscope, regardless of its sophistication, contains the following component blocks (Fig. 11-1):

1. Display device (the CRT, descended from Crookes tube, still serves this function).
2. Vertical amplifier (including probe or transducer to obtain an electrical signal).
3. Time base.
4. Horizontal amplifier.
5. Trigger or sync circuit (to start each sweep at a desired point in time for the signal to be displayed).
6. Gate amplifier (to turn CRT intensity *on* while the beam is swept horizontally; *off* at other times).
7. Power supplies.

Laboratory oscilloscopes may be classified many ways. Usually the distinctions are based either on frequency-response capability or on CRT characteristics. Thus there are *low-frequency* oscilloscopes (to ≈ 10 MHz for vertical amplifier response), *high-frequency* oscilloscopes (sometimes capable of capturing and displaying single-shot phenomena of less than 1-nsec rise time), and *sampling* oscilloscopes (which reconstruct very high frequency—to 18 GHz—repetitive waveforms on a dot-sample basis). There are *standard refreshed* phosphor oscilloscopes and *storage* oscilloscopes, depending upon the type of CRT used.

11-2 Oscilloscope CRTs

The CRT is the most distinctive component of an oscilloscope since it is the obvious output or display portion of the instrument. In recent years, much research has been conducted to obtain a better display means than a CRT, and several alternatives have been developed, including electroluminescent panels, solid-state light-emitting arrays of gallium arsenide diodes, and plasma cells. For the present and foreseeable future, however, it seems clear that today's refined versions of the original Crookes tube will continue to dominate the displays found in oscilloscopes, because of advantages in cost, brightness, and speed of response.

Cathode ray tubes may be classified in several important ways. By the number of independent electron beams a tube is classed as *single-beam*, *dual-beam*, or *multibeam*. There are differences in beam-deflection method, resulting in *electromagnetic* CRTs and *electrostatic* CRTs. There are distinctions within electrostatic tubes based upon deflection-plate design, giving *parallel-plate*, *bent-plate*, *segmented-plate*, and *distributed-plate* CRTs. The absence or presence of beam acceleration between the deflection plates and the phosphor target gives rise to further differences: *monoaccelerator* and *postaccelerator* CRTs.

Target-decay characteristics are the most apparent distinctions between CRTs for the typical user. The decay and refresh characteristics required to maintain a flicker-free display vary not only as a function of phosphor type, but according to design differences between *standard* CRTs (requiring continual phosphor refresh), *storage* CRTs, and *variable-persistence* CRTs.

Lastly, distinctions are occasionally noted between *internal-graticule* and *external-graticule* CRTs, between *grid blanking* and *deflection blanking* for CRT intensity, and *electrostatic* or *electromagnetic focusing* methods. Gradations in such parameters as *spot size*, *light output*, *contrast ratio*, *writing speed*, *linearity*, and *deflection defocusing* are also significant in many applications [4, 5, 6, 7].

The operation of a CRT may be described by the five regions illustrated in Fig. 11-2*a*. The regions are the *beam-generation* area,where the beam is generated; the *beam-focus* section, where the beam is focused; the *beam-deflection* region, where the beam is positioned to the desired *xy* coordinates; the *beam-postacceleration* section, where space potential controls the beam velocity; and the *beam-target*, or *screen*, region, where the display characteristics are observed.

Beam Generation. The beam-generation region is analogous to a triode: a cathode supplying electrons, a grid controlling their rate of emission, and an anode collecting them. In a CRT, there is a small hole or aperture in both the grid and the anode, permitting a narrow beam to emerge from

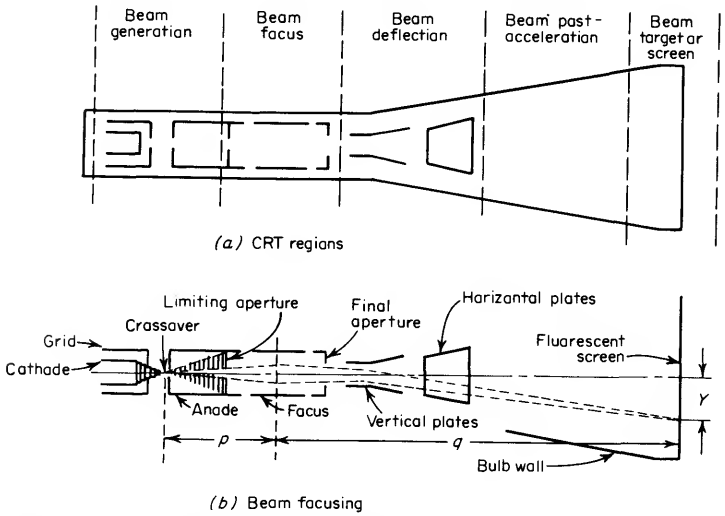


FIG 11-2 Cathode-ray-tube internal structure.

the anode. The cathode potential is very negative, often several thousand volts, with respect to the anode, so that the beam emerging from the anode has attained considerable energy [8].

The intensity of the beam at the target is most conveniently controlled by adjusting the bias voltage of the grid; this grid is usually driven by the z-axis amplifier in a grid-blanking CRT. It is also possible to make a double anode aperture, with intermediate deflection plates that may be used to deflect the beam to miss the second aperture. Because the beam is not stopped at the source but merely deflected enough that it is not seen at the screen, this method is termed *deflection blanking*. One advantage of deflection blanking is that power-supply regulation is not nearly as critical as in a grid-blanking scope; the fact that the cathode is always emitting large quantities of electrons, with some consequent cathode life reduction, must be weighed against the power-supply cost saving.

Beam Focus. The beam-focusing region contains the electrodes intended to cause the beam pattern to converge as a small round dot on the target. The two usual electrodes are the *focus* and *astigmatism* electrodes. Although these two electrodes interact to some extent, in general it is true that the focus control adjusts the CRT lens action for the smallest dot at the target, while the astigmatism control is used to make the dot as nearly round as possible, both at screen center and at the perimeter. The astigmatism electrode thus may provide some correction of focus lens

aberrations. Front-panel FOCUS and ASTIGMATISM controls are typically provided for the operator to optimize both settings. Occasionally tubes are made with provision for electromagnetic field focusing with external magnetic coils for focus control. These are of merit in small-spot-size, large-screen tube designs where internal electrodes give patterns that create edge distortion. More commonly, external coils are used to adjust the x-axis and y-axis deflection patterns. The TRACE ALIGN control allows adjustment of an x-axis deflection to a horizontal graticule line on the CRT screen. Perpendicularity of the y axis to the x axis may be adjusted if a magnetic coil is provided to adjust the y-axis ORTHOGONALITY.

Electrons emitted from the cathode surface are subject to a lens action at the grid aperture which leads to a minimum beam cross-sectional area called *crossover* (Fig. 11-2*b*). The beam then reexpands within the anode structure, and in fact expands until reaching the focus lens, which serves to focus the beam so that it converges to the smallest dot at the screen. The relative lengths of p and q determine the image-to-object ratio q/p. This ratio times the beam-crossover spot size gives a rough index of the minimum screen *spot size* obtainable on a CRT.

Beam Deflection. Many possible combinations of cost and performance trade-offs are available in the beam-deflection region, which accounts for the great variety of CRTs that differ only in deflection method. Magnetic deflection allows a wider beam-deflection angle than does electrostatic deflection. Moreover, when the full-screen beam-deflection bandwidth desired is less than 20 kHz, the electromagnetic deflection system (amplifier and CRT) has a substantial cost advantage over an electrostatic system. Thus television sets, many medical monitors, and some oscilloscopes rely upon CRTs with electromagnetic deflection.

Magnetic deflection is accomplished by changing magnetic field patterns, and this is done by changing current levels in an inductor. At high frequencies, inductors with few turns are necessary to obtain fast current changes, and larger currents are required to obtain the required field strength. Consequently, above repetition rates of 20 kHz, large power dissipations are required to obtain full-scan displays. Electrostatic deflection, involving voltage charging of capacitive plates, is capable of speeds several orders of magnitude higher for a comparable amplifier cost (but not CRT cost). Since even inexpensive industrial oscilloscopes generally are capable of displaying 500 kHz or more on the vertical axis (and often on the horizontal axis as well), it is not surprising that most CRTs in oscilloscopes use electrostatic deflection.

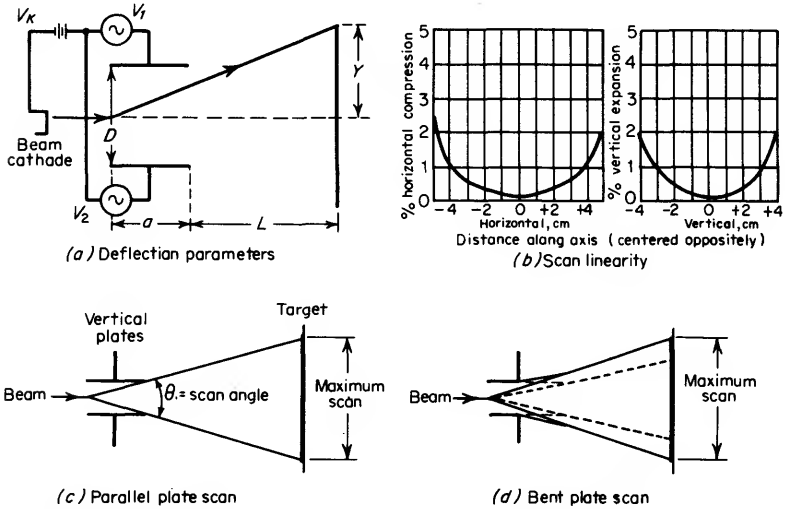
Parameters of significance in an electrostatic tube are *deflection* or *sensitivity factor* (the number of volts of differential potential required to move the beam one screen division up or down), *linearity* (whether or not the deflection sensitivity is the same at center screen as it is near the edges),

and *scan area* (the number of screen divisions which can be scanned—full-screen scan is often achieved—before the beam intercepts the deflection plate itself). Figure 11-3 illustrates these three parameters, and it may be seen from the first diagram of the figure that the deflection Y is directly proportional to the deflection voltage $V_1 - V_2$, length L, and length a, and inversely to the distance between the plates, D, and the beam potential V_k. Thus the electrostatic-tube deflection factor DF_{es} may be written as

$$DF_{es} = \frac{V_1 - V_2}{Y} = \frac{K_1 V_k D}{aL} \tag{11-2-1}$$

where K_1 is a constant determined by the postaccelerator field.

From the foregoing, it is apparent that an electrostatic CRT designed for minimum deflection factor (greatest sensitivity) would have relatively long plates spaced closely together, a long tube, and a low beam potential. On the other hand, an electromagnetic-CRT deflection factor is not affected as greatly by increasing beam potential, which means that, in general, magnetic tubes can provide better brightness and resolution since higher beam energy can be attained without losing as much sensitivity. Shorter tubes can be built with magnetic deflection than with electrostatic designs for the same display area since there are no deflection plates inside the tube to limit the scan angle and because aberrations are fewer with magnetic deflection.



(a) Deflection parameters

(b) Scan linearity

(c) Parallel plate scan

(d) Bent plate scan

FIG 11-3 Cathode-ray-tube parameters.

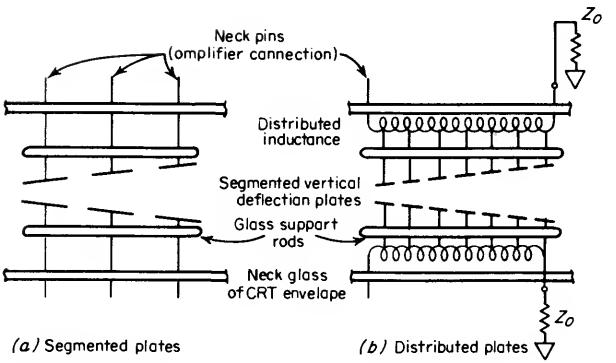


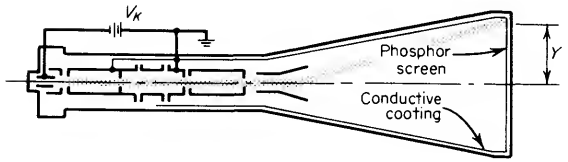
FIG 11-4 Cathode-ray-tube deflection-plate designs.

Even for electrostatic tubes, the scan angle may be increased by using *bent-plate* or *curved-plate* designs instead of *parallel-plate* structures (Fig. 11-3*c* and *d*). Note that this allows either a shorter tube for the same scan, or a larger scan for the same tube length. Other electrostatic deflection-plate designs include *segmented plates* and *distributed plates* (Fig. 11-4), which reduce capacitive loading effects for extremely high-speed deflection (full scan-angle deflections in times less than 5 nsec).

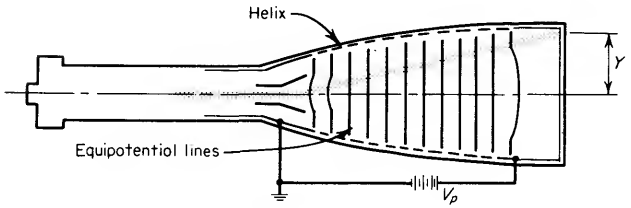
Beam Postacceleration. The fourth region, the beam-postacceleration region, is of great significance to *writing speed* or beam brightness at fast deflection rates. Monoaccelerator designs are intended for low-frequency applications; the name suggests that no accelerating field is found in the postaccelerator region (the original accelerating field is the potential between the cathode and the deflection plates). Postaccelerator designs (often termed *PDA tubes*, for *postdeflection acceleration*) are generally classified in present instrumentation as either grid or shaped-field tubes: frame-grid or radial-field mesh structures combined with a constant field potential between the grid and the screen are the first type, and a spiral, or helix, with a variable field potential represents the second type. They all use a large positive bias voltage V_p beyond the deflection plates to increase the beam velocity and energy to obtain a brighter dot on the target.

The monoaccelerator CRT of Fig. 11-5*a* performs satisfactorily in oscilloscopes for viewing relatively low frequency signals (below 10 MHz). In such a tube, the beam is given all its acceleration before it passes through the deflection plates and arrives at the screen with sufficient velocity to present a bright display for relatively low writing speeds.

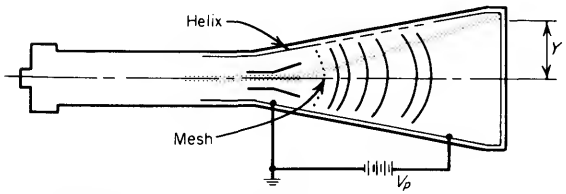
In electrostatic CRTs, the voltage V_k from cathode to deflection plate is usually less than 4 kV in order to keep deflection sensitivity high. Consequently, at higher writing speeds, the beam must be accelerated



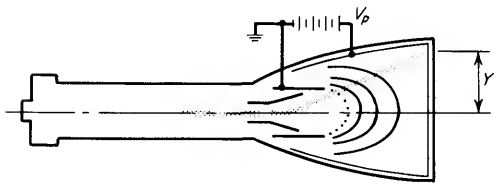
(a) Monooccelerator



(b) Shaped-field postaccelerotor



(c) Rodiol-field mesh postoccelerotor



(d) Exponsion mesh postoccelerotor

FIG 11-5 Cathode-ray-tube beam-acceleration techniques.

after deflection to produce a bright display. A shaped-field post-accelerator CRT uses a resistive spiral inside the tube envelope upon which is impressed an accelerating voltage of the order of 10 kV, Fig. 11-5*b*. The accelerating field bends the beam toward the axis and thus demagnifies the waveform displayed or, in effect, decreases deflection sensitivity. To achieve a display of equivalent size to that of the monoaccelerator tube, the helix postaccelerator tube must be made longer. A flat frame-grid

structure inserted between the deflection plates and the electrode at V_p potential will achieve the same performance as the shaped-field tube.

If a spherical mesh is inserted into the helix tube, the accelerating field is shaped to avoid the compression of the display in the conventional postaccelerator CRT, Fig. 11-5*c*. This is called a radial-field mesh CRT, and its sensitivity and display size equal that of a monoaccelerator tube of the same length.

Carried one step further, the helix can be eliminated and the mesh shaped into a configuration that will expand or magnify the display. The result is a short tube that behaves like a long one, Fig. 11-5*d*. This CRT, called a *high-expansion* (or *high-magnification*) *mesh tube*, achieves an increased deflection sensitivity and scan angle. With such expansion-mesh techniques, electrostatic deflection large-screen CRTs have been made with scan angles approaching 90° (Fig. 11-6), which compares well with the present typical 110° in magnetic deflection tubes. By contrast, any of the first three types of electrostatic tubes discussed are limited to a scan angle of about 35°. Expansion-mesh CRTs expand not only sensitivity and scan angle, but also spot size proportionately.

The electrodes of the first three regions are termed the *gun* structure, which is assembled as a complete entity and inserted into the *neck* of the glass envelope. The target and postacceleration regions are commonly built into the funnel of the envelope. The two parts are then fused together, and the entire tube is evacuated and sealed at the base (cathode end).. Typically, all connections to internal electrodes are made through

FIG 11-6 A 14-in. diagonal expansion-mesh electrostatic-deflection CRT. (*Hewlett-Packard Company.*)

the base plug for monoaccelerator tubes; for postaccelerator tubes intended for higher-frequency deflection rates, the deflection-plate leads are brought out through the neck glass on *neck pins* to reduce circuit capacitance (Fig. 11-4), and the postaccelerator voltage connection is made on the funnel to avoid arcing. Additionally, storage tubes usually have connections on the edges of the envelope in the target region.

11-3 Cathode-ray-tube Display-screen Characteristics

The display-target screen, or region, has to do with the display provided for the user [9, 10]. The front face of the CRT, called the *faceplate*, is the picture screen of the oscilloscope. This face is usually a flat surface for tubes with display area *windows* of 10 × 10 cm or less, and is a slightly convex surface for larger tubes.

Phosphor Characteristics. Phosphor is the usual readout material on the target; it has the capability of converting electrical energy into light energy. Two distinct phenomena occur when a phosphor is bombarded with a high-energy electron beam. When the beam hits the phosphor, a *fluorescence*, or light emission, is observed. When the excitation beam is removed, a *phosphorescence* remains for some time and indicates where the phosphor had been stimulated into light emission. Standard CRTs use *refreshed phosphor* targets, so named because the phosphor must be restimulated or refreshed before the phosphorescence has decayed in order to avoid a blinking or flickering display. Storage CRTs by contrast do not require continual refresh to maintain a displayed trace.

Phosphors vary greatly in fluorescence excitation times and colors as well as the phosphorescence decay times and colors (Table 11-1). Phosphors are classified as *short-persistence* (decay of intensity to $1/e$ of the excitation level in less than 1 msec), *medium-persistence* (decay in less than 2 sec), and *long-persistence* (decay may take minutes or more). The high repetition rates required for most waveform displays make a long-persistence phosphor unnecessary and even undesirable because of the trace afterglow; consequently most scopes use a short-persistence phosphor such as P1, P2, P11, or P31. Medical displays require a longer phosphor decay because of the slow repetition rate of human physiologic functions such as pulse pressure and heart rate; phosphors such as P7 and P39 are of value in such applications. Very slow displays (such as radar) require either long-persistence CRT phosphors or storage techniques to maintain an adequately flicker-free picture; P19, P26, and P33 phosphors are used in these applications.

Phosphor color is occasionally important; the human eye tends to peak in the yellow-to-green region ($\approx$5,500 Å), which means that a P2 or P31 phosphor will appear brighter to the eye under the same CRT conditions

TABLE 11-1

Phosphor type	Color		Persistence	Relative luminance	Relative writing speed	Display application
	Fluorescence	Phosphorescence				
P1	Yellow-green	Yellow-green	Medium	45	35	Scopes, radar
P2	Blue-green	Green	Medium	60	70	Scopes
P4	White	White	Medium to medium short	50	75	Black & white television
P7	Blue-white	Yellow-green	Blue–med. short Yellow–long	45	95	Radar, medical
P11	Blue-violet	Blue	Medium short	25	100	Photographic recording
P15	Blue-green	Blue-green	Visible–short UV–very short	15	25	Flying spot scanners for TV
P16	Blue-violet	Blue-violet	Very short	0.1	25	Flying spot scanners for TV
P18	White	White	Medium	18	35	Low-frame rase television
P19	Orange	Orange	Long	25	3	Radar
P22	3-color dot pattern, red, blue, green		Medium	....	...	Color television
P26	Orange	Orange	Very long	17	3	Radar
P28	Yellow-green	Yellow-green	Long	50	50	Radar, medical
P31	Green	Green	Medium short	100	75	Scopes
P33	Orange	Orange	Very long	20	7	Radar
P39	Green	Green	Medium to medium long	50	40	Computer graphics

than does a P11, P16, or P19. Camera film is often sensitive to ultraviolet or blue; hence for high-writing-speed photography, a P11 or P16 is well suited. Monochromatic television using P4 phosphor (black and white) is certainly more pleasing to the home viewer than a P31 green.

The kinetic energy of the electron beam is converted into both light and heat energy when it hits the target. The heat gives rise to *phosphor burn* on occasion, which is damaging and sometimes destructive. Excessive beam current density for a period of time at one spot can degrade the light output of a phosphor, and in extreme cases can burn a spot in the phosphor which amounts to complete phosphor destruction. Phosphors may be classified according to burn resistance as low (P19, P26, P33), medium (P1, P2, P4, P7, P11), and high (P15, P31). Thus, the typical choice of P31 for oscilloscope CRT phosphor is partially based on color (maximum eye response), short persistence (to avoid multiple-image

displays when the image moves rapidly), and high burn resistance (to avoid accidental damage by the operator), as well as on its high luminance level and high writing speed.

Aluminizing a CRT (depositing a thin layer of aluminum on the nonviewed side of the phosphor) accomplishes three goals. Its original function was to avoid buildup of charges on the phosphor, which tends to slow down the electrons and limit brightness. Also, aluminizing serves to reduce light scatter (emission of light in all directions) when the beam hits the phosphor. With the aluminized layer, the light emitted back into the tube is reflected again toward the viewer and thus increases the brightness, as shown in Fig. 11-7. Finally, the aluminum layer acts as a heat sink for the phosphor and thus materially reduces the danger of phosphor burning. Since penetration of the aluminized layer requires considerable beam energy, the efficiency of such CRTs is low at low acceleration potentials, although it is significantly higher at higher potentials.

Light filters are used to increase contrast or to accentuate either a short- or long-persistence color component of a phosphor. Thus, a smoke-gray filter will darken a light background and give an apparent brightness increase to a green phosphorescent area. On a P7 phosphor, an amber filter will eliminate the short-term blue component and increase the longer-persistence orange component, while a blue filter will do the reverse. Clear or colored plastic filters (such as Plexiglas acrylic plastic or Lucite acrylic resin) may be used for such purposes and in fact are often supplied with oscilloscopes.

Filters are also used to reduce glare from ambient lighting. Black wire-mesh filters restrict the viewing angle and thus minimize reflections from oblique light sources (such as overhead lights) while also enhancing the contrast ratio. Polarized filters are occasionally used in high-glare situations (such as hospital surgery rooms).

Spot Size and Luminance. Spot size is not only determined by crossover-spot size and image-to-object ratio, as earlier indicated, but also

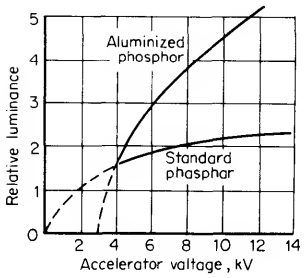


FIG 11-7 Cathode-ray-tube luminance.

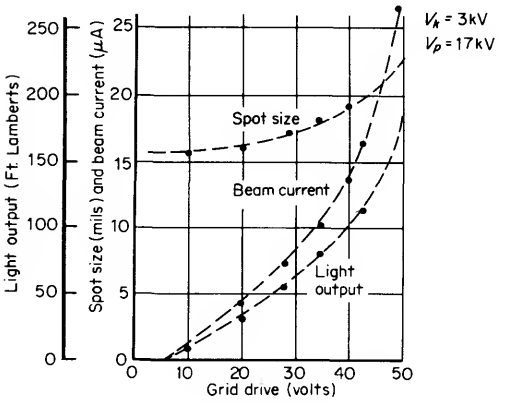


FIG 11-8 Cathode-ray-tube spot size.

by beam current density changing the crossover-spot size, other space charge effects, and deflection defocusing. As shown in Fig. 11-8, for a z-axis grid bias below about 20 V, the crossover-spot size is relatively constant for postaccelerator mesh tubes. Between about 20 and 40 V of z-axis unblanking, the spot size changes in a fairly linear fashion, which indicates that higher current demands have changed the crossover-spot size. Above 40 V, space-charge effect (the mutual repelling action of electrons in the beam) begins to cause more rapid increases in spot size than might otherwise be expected and no lens correction can be satisfactorily achieved. Deflection defocusing increases spot size at the edges of the CRT in comparison with the center, primarily because of beam distortion as the beam approaches a deflection plate.

Cathode-ray-tube brightness or intensity is a function of beam energy and phosphor. Beam energy is the product of beam current density times accelerating potential times writing time and is essentially linearly related to light-output brightness (Fig. 11-8). Thus monoaccelerator CRTs with low beam current may still give a bright trace at slow sweep speeds. Conversely a single-shot high-speed transient waveform may only be displayed on a CRT with high beam current and high accelerating voltage. Relative luminance of phosphors (Table 11-1) is a measure of the relative brightness (as seen by the eye) to be expected from each phosphor under the same operating conditions of the CRT.

Writing Speed. Photographic writing speed is a measure of the fastest deflection rate of a single beam trace (single-shot waveform display) that is just barely visible on film. It is affected not only by CRT parameters such as density of beam current, accelerating voltage, and phosphor, but also by parameters such as light filters, camera lenses, and film speed

FIG 11-9 Several electrostatic-deflection CRT designs. (*Hewlett-Packard Company.*)

and sensitivity. Cathode-ray-tube writing speed is typically given in centimeters per microsecond. If, for example, the scope sweep time is 10 nsec/cm and a single-shot ramp 5 cm high and 50 nsec wide is to be photographed, the writing speed must be at least 140 cm/μsec. The relative writing speed of a phosphor does not quite correlate with the relative luminance, because film is more sensitive to the blue portions of the spectrum than is the eye. Otherwise, factors affecting luminance (such as beam energy, filters, spot size) also affect writing speed in a similar fashion.

Graticules. Graticules are scale markings on or near the CRT screen to aid the user in display analysis by calibrating beam deflection. The number of scale divisions multiplied by the appropriate switch settings gives the amplitude difference and time duration between any two points on an observed waveform. Three kinds of graticule systems are in use: *external graticule*, *internal graticule*, and *projected graticule*. The external graticule, scribed on a piece of Plexiglas acrylic plastic, is easily changed to accommodate several kinds of measurements. It does, however, suffer from parallax since the graticule is not in the same plane as the phosphor. The internal graticule overcomes this disadvantage, as it is deposited onto the internal surface of the CRT faceplate, effectively on the same surface as the phosphor. It is not changeable, and it is difficult to illuminate for photography without either an ultraviolet light source within a camera or special flood illumination provisions in the scope. Projected graticules are provided with some cameras, and these allow the greatest flexibility in graticule patterns and even nomenclature inclusion.

Graticule sizes and patterns vary greatly (Fig. 11-9). For oscilloscope displays, round CRTs (usually 5 in. in diameter) were standard for a number of years, and these allowed displays ranging in size from 4 × 10 cm to 10 × 10 cm (with notched corners). More recently, rectangular CRTs have been introduced, with displays from 4.8 × 8.0 cm to 8 × 10 cm currently available in scopes with wide-ranging capabilities. Larger-

screen electrostatic CRTs up to 8 × 10 in. in size and larger graticules are available for many oscilloscope measurements.

11-4 CRT Storage-target Characteristics

Storage targets can be distinguished from standard phosphor targets just discussed because they have the ability to store or retain a waveform pattern for a time, independent of phosphor persistence. Two storage techniques are used in oscilloscope CRTs: *mesh storage* and *phosphor storage*. A mesh-storage CRT uses a dielectric material deposited on a storage mesh as the storage target. This mesh is placed between the deflection plates and the standard phosphor target in the CRT. The *writing beam*, the focused electron beam of the standard CRT, charges the dielectric material positively where hit. The storage target is then bombarded with low-velocity electrons from a *flood gun*, and the positively charged areas of the storage target allow these electrons to pass through to the standard phosphor target and thereby reproduce the stored image for the viewer to observe. Thus the mesh-storage CRT has both a storage target and a phosphor display target. The phosphor-storage CRT uses a thin layer of phosphor to serve as both the storage element and the display element. Both kinds of storage targets are capable of use as *scan converters*, but the mesh-storage tube is uniquely suited to intensity-modulation and variable-persistence operation, while the phosphor-storage tube is more capable of giving display regions for storage and nonstorage simultaneously [11, 12, 13, 14].

Mesh Storage. A mesh-storage CRT contains a dielectric material deposited on a storage mesh, a collector mesh, a flood gun, and a collimator, in addition to all the elements of a standard CRT (Fig. 11-10). The storage target, a thin deposition of a dielectric material such as magnesium fluoride on the storage mesh, makes use of a property known as

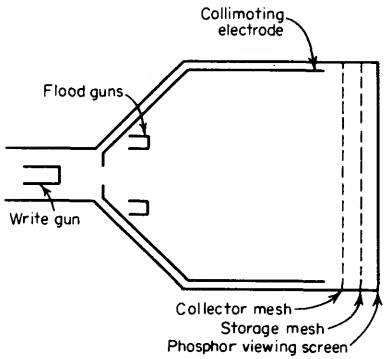


FIG 11-10 Mesh-storage CRT structure.

secondary emission. This characteristic, illustrated in Fig. 11-11, is common to most materials when bombarded by electrons of sufficient energy. Between first and second crossover, more electrons are emitted than are absorbed by the material, and a net positive charge results. Below first crossover a net negative charge results, since impinging electrons do not have sufficient energy to force an equal number to be emitted. In the presence of a flood gun biased at ground and a collector mesh biased at V_c (volts), two stable points of operation may be found for the storage surface (Fig. 11-11*b*). Note that these points bracket the first crossover region, which is an unstable point.

In order to store a trace, assume that the storage surface is uniformly charged to the lower stable point, and the *write gun* (the beam-emission gun of the standard CRT) is biased well beyond first crossover with respect to the storage-mesh potential. Thus, writing-beam electrons will hit the storage target with energy eV_k as in Fig. 11-11*a*, where the secondary-emission ratio is much greater than unity. Those areas of the storage surface hit by the deflecting beam will lose electrons, which are collected either by the collector mesh or the display phosphor target. Thus the write-beam deflection pattern is traced on the storage surface as a positive charge pattern. Since the insulation of the dielectric material is adequate to prevent charge migration for a considerable length of time, the pattern is effectively stored.

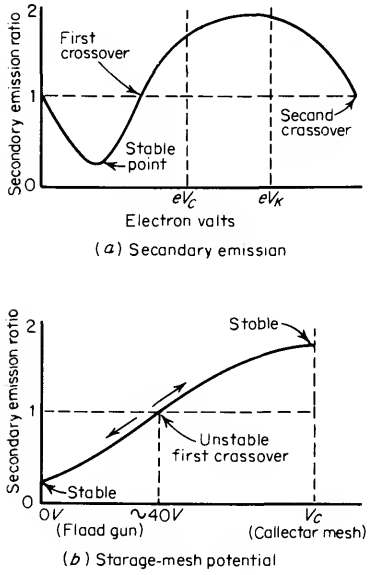


FIG 11-11 Secondary-emission characteristics.

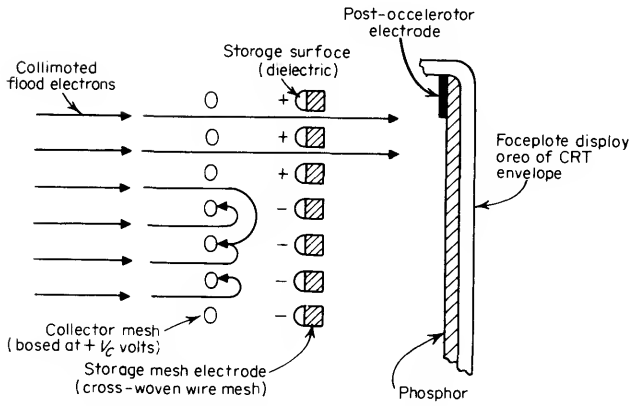


FIG 11-12 Display of the stored-charge pattern on a mesh-storage CRT.

To view the stored trace, a flood gun is used when the write gun is turned off. The flood gun, biased very near the storage-mesh potential, emits a great flood of electrons, which migrate toward the collector mesh since it is biased slightly more positive than the deflection region. The collimator, a conductive coating on the CRT envelope with an applied potential, helps to align the flood electrons so that they approach the storage target perpendicularly. When the electrons penetrate beyond the collector mesh, they encounter either a positively charged region on the storage surface or the negatively charged regions where no trace has been stored (Fig. 11-12). The positively charged areas allow the electrons to pass through to the postaccelerator region and the display target phosphor. The negatively charged regions repel the flood electrons back to the collector mesh. Thus the charge pattern on the storage surface appears reproduced on the CRT display phosphor just as though it were being traced with a deflected beam.

The stored pattern eventually degrades, primarily because ions generated by flood-gun electrons charge other regions of the storage surface and the entire display consequently appears to be written. This is called *fading positive*. A typical mesh-storage CRT will store a trace for an hour or more, and the trace may be displayed at bright intensity for at least a minute.

To erase the storage surface of stored traces, a momentary-contact ERASE button is provided. This biases the storage-mesh potential at the same level as the collector mesh, and because of capacitive coupling between the storage mesh and the storage surface, the surface is taken to a potential well above the first crossover voltage. Flood electrons now

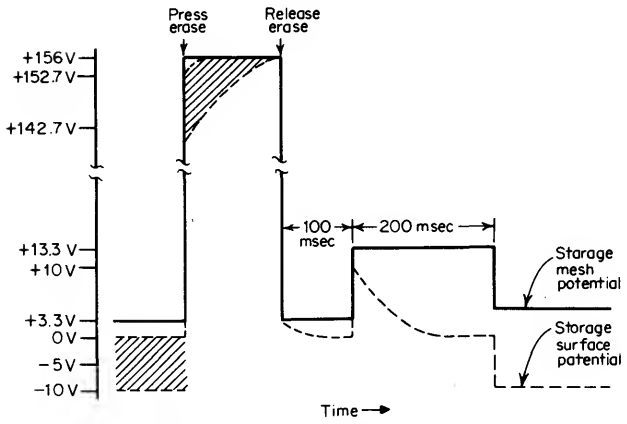


FIG 11-13 Storage-mesh potentials for a mesh-storage CRT during erase cycle.

quickly charge all storage surface areas to the collector-mesh potential. The storage surface cannot exceed the collector-mesh potential because any further secondary emission is repelled back to the storage surface. Note that this is the upper stable point of Fig. 11-11*b*. Figure 11-13 illustrates the bias voltages for both the storage surface and the storage mesh during the complete erase cycle for a CRT produced at present.

When the ERASE button is released, the storage mesh returns to the level it had before the button was pressed (+3.3 V for the tube in Fig. 11-13). This level is held electronically for 100 msec, during which time the flood electrons land on the storage surface and it is charged back to 0 V (the flood-gun cathode voltage, and the lower stable point of the storage surface). Then the storage mesh is raised to +13.3 V, and the storage surface goes to +10 V by capacitive coupling. Since this is still well below first crossover for flood-gun electrons, the storage surface again decays to 0 V in a time less than 200 msec. After 200 msec, the storage mesh is again returned to +3.3 V, and this time the storage surface goes to −10 V, which is the desired condition for flood-gun electron repulsion, as shown in Fig. 11-12.

Note that a writing beam, in order to be stored, must have sufficient time to charge the written storage surface to a level such that flood-gun electrons are permitted to pass on through to the display target. This occurs in the above CRT at about −5 V, with greater contrast attained as the surface region nears ground potential. The time required for a beam to charge a surface region to this threshold level largely determines the writing speed of a storage target. If the storage surface were only taken to about −6 V instead of −10 V, the writing speed of the storage

target could be greatly increased. A MAXIMUM WRITE mode is often provided to do just this, and for the tube cited, the writing-speed improvement is about 50 times, going from 20 divisions/msec to 1 division/μsec. The disadvantages of operating in this mode include reduced storage time, reduced contrast ratio, and nonuniformity of display since the threshold level may vary slightly across the entire storage surface.

The mesh-storage tube described above is capable of displaying several different levels of intensity, known as *gray scales* or halftones. Since the storage surface has a threshold level and the increase of charge beyond that level allows more flood electrons through to the phosphor target, a gradient of flood-gun illumination may be achieved by varying the charge levels on the storage surface. This may be done by either varying the "time dwell" of the beam on the storage surface or by varying the writing-beam current with a constant time dwell. The useful levels may approach as many as the 10 required for good television-picture presentation, but this requires a highly uniform storage surface. Four levels, roughly representing −10 (full *off*), −5, −2.5, and 0 V (full brightness) on different regions of the surface, are practically achievable.

Phosphor Storage. Although work on storage targets with secondary-emission properties has been conducted since the start of the century, a storage-target CRT design based on the dielectric mesh target was first constructed in 1947 by Dr. Andrew Haeff. Some years later, Robert Anderson developed the bistable phosphor storage target by using similar principles of secondary emission. These two approaches comprise nearly all commercial storage-CRT construction today. Significant improvements in both types are regularly forthcoming.

In the *bistable storage tube*, the same material is used for both the storage target and display phosphor (Fig. 11-14). The material used is a *P*1 phosphor doped for good secondary-emission characteristics. A necessary condition for such a target is that boundary migration of stored charge must be eliminated; scattered phosphor particles achieve this condition provided that the deposition is shallow enough that the surface is not

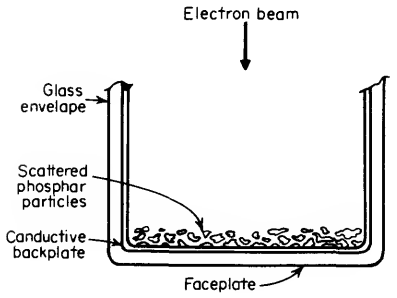


FIG 11-14 Target structure of a phosphor-storage CRT.

electrically continuous. The layer may be more than one particle thick, which allows a viewing phosphor of continuous nature, but there is a thickness threshold beyond which no storage is possible. The controlling electrode for a bistable phosphor is the *conductive backplate*, a transparent metal film deposited on the inside surface of the faceplate before the phosphor is deposited. A wide range of operating voltages on this electrode (about 100 to 200 V) gives a stable storage characteristic; voltages below 100 V will uniformly erase the target, and voltages above 200 V will uniformly write the target.

The bistable nature of the storage on the phosphor means that a trace is either stored or it is not, and brightness is thus *on* or *off*. Halftones are therefore not possible, nor is variable persistence, although simulated effects may be produced [15]. The bistable tube has been manufactured in a *split-screen* version by depositing two independent conductive backplates, one covering the top half of the CRT viewing area and the other covering the bottom half. Thus, by operating the top backplate at ≈ 150 V for the flood guns, and the bottom half at ≈ 50 V, the CRT is a storage tube on the top half and a standard refreshed phosphor display on the bottom half.

Both phosphor-storage and mesh-storage tubes are susceptible to burning; a typical burn appears as a stored trace that cannot be erased. The mesh-storage tube may have trapped charge below the storage surface which flood electrons during the erase cycle are incapable of reaching; this kind of burn may usually be erased with a continuous erase mode for several hours. Either tube may suffer target burn much like phosphor burn previously discussed; this is destructive in every case. The phosphor-storage tube, since it uses a very thin doped phosphor coating, is susceptible to light output reduction with use. This aging characteristic results in a relatively short CRT life.

Variable Persistence. The mesh-storage technique is capable of using a complex erasure pulse to vary the effective persistence seen on the screen. This is really a storage mode with continuous electrical control of the duration of storage. A phosphor-storage tube is capable of erasure electronically at the end of every sweep, but continuous persistence control is denied it because of the bistable nature.

Variable persistence allows continuous control of the persistence from about 200 msec to several minutes and thus allows the trace persistence to be set for a small trace-decay tail to indicate recent past history of a slow-moving dot (radar use), to show five or more sequential traces (to observe medical anomalies), or to have the previous trace fade just as a new one replaces it [16].

Figure 11-15 shows the erase pulse used to obtain variable persistence in the mesh-storage CRT discussed previously. As mentioned before,

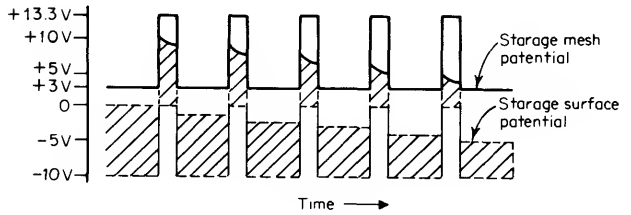


FIG 11-15 Storage-mesh potentials for variable-persistence operation.

the storage-mesh potential is held at +3.3 V, and the unwritten storage surface is at −10 V. Charged areas are near 0 V if written long enough. If a +10 V pulse is applied to the storage mesh, the unwritten areas of the storage surface will be about 0 V, while the written areas are near +10 V. The written areas, still below first crossover, will now attract flood-gun electrons and decay to a lower voltage. When the pulse ends, the unwritten areas return to −10 V and the positive regions are slightly below ground. A number of pulses will ultimately drop the written areas below the −5-V viewing threshold, and the "persistence" ceases. The time required to erase a trace in this manner is controlled by varying the duty cycle of the pulses applied to the storage mesh.

Scan Conversion. Scan conversion is a technique for converting an analog electrical signal (such as that at the deflection plates of a CRT) to a digital time-coded signal for use in radar displays, computer analysis, television video, or long-distance data transmission. The storage target of either type of CRT just discussed is adaptable to such scan conversion.

Along with, or instead of, a phosphor display target, a metal-electrode target is provided behind the storage target. The flood gun is replaced by a focused *interrogate* gun complete with deflection plates. Once a trace is stored, in a fashion similar to that of any storage target, the interrogating beam may be swept in a raster-scan manner (much like television) across the target. Every time a positive-charge point is encountered, a pulse of current output is sensed on the metal "display" target. Thus the output becomes a small current pulse that is time sequenced according to the *xy* coordinates of the raster-scan beam location. This type of signal lends itself well to many emerging industry needs, such as digital communications [17, 18].

11-5 General-purpose Oscilloscopes

The general-purpose oscilloscope incorporates an electrostatically deflected CRT with a standard refresh phosphor and real-time display

circuitry. In a real-time display, the CRT beam deflection faithfully traces the signal across the screen as the event occurs. For repetitive signals, the beam traces the pattern over and over, and a bright waveform is displayed. A *sampling oscilloscope* (Sec. 11-6) is also able to display a repetitive waveform, but it does not do it in real time. Instead, one dot is displayed on the screen every time that a signal cycle occurs, and then the dot sample point is advanced slightly each cycle, so that a series of advancing dots reconstructs the signal much as an optical stroboscope reconstructs a picture. A signal that occurs only once, such as an impact test or an explosion, must be displayed on a real-time oscilloscope, for only one dot would be displayed on a sampling oscilloscope.

An electromagnetically deflected CRT may be used with real-time display circuitry, but the beam deflection speed is typically limited to frequencies under 1 MHz. This is too slow for observations of many electrical phenomena, although adequate for most physiologic, mechanical, and chemical measurements.

Storage CRTs are usually used with real-time display circuitry; their chief value is in either storing single-occurrence events or providing displays at low-frequency repetition rates without flicker.

All general-purpose oscilloscopes use similar techniques for vertical amplifiers, time bases, and trigger generators that are fundamentally different from the techniques of sampling circuitry. The most pertinent specifications among different general-purpose oscilloscopes are *number of channels*, *deflection factor*, *common-mode rejection* (CMR), and *rise time* of the vertical amplifier system; the *sweep speeds* and *number of display modes* of the time base; and the *trigger circuit capability* [19, 20].

Bandwidth and Rise Time. Just as electronic circuits in general have historically been rated in terms of bandwidth even though pulse-transient analysis is often more meaningful, so the oscilloscope is ordinarily described by a rated bandwidth: from dc to X Hz. This means that at X Hz, the vertical amplifier is down no more than 3 dB from its dc or zero frequency gain. Since an oscilloscope is an instrument often intended for pulse analysis, it is more directly specified in capability by its limiting rise time or beam deflection rate. Rise time is defined as the time required for a pulse to rise from 10 to 90 percent of its final value.

Most oscilloscope amplifiers are designed for minimal pulse distortion (in contrast to a maximally flat frequency spectrum), and to the extent that a pulse response is free from overshoot distortion (<2 percent overshoot or ringing), a convenient relationship between bandwidth B and rise time T_r exists:

$$B = \frac{K}{T_r} \qquad K \approx 0.35 \tag{11-5-1}$$

where bandwidth is in megahertz, and T_r is in microseconds. Thus a 35-MHz oscilloscope is capable generally of a 10-nsec rise time. Since different amplifier response characteristics may cause the constant given above to vary between 0.3 and 0.5 (0.35 is characteristic of a single-pole RC roll-off), rise time is usually the more significant specification for a time-domain response, while bandwidth is more meaningful for the frequency-spectrum measurement.

For measurements of signal rise times approaching that of the oscilloscope itself, the error caused by the instrument rise time should be compensated for. This may be done with the following equation for overall rise time:

$$T_{rd} \approx [(T_{rs})^2 + (T_{ro})^2]^{1/2} \tag{11-5-2}$$

where T_{rd} is the overall displayed rise time, T_{rs} is the signal rise time, and T_{ro} is the oscilloscope rise time. This may be rewritten to solve for the signal rise time

$$T_{rs} \approx [(T_{rd})^2 - (T_{ro})^2]^{1/2} \tag{11-5-3}$$

For example, if an oscilloscope has a specified rise time of 10 nsec, and the displayed rise time is 18 nsec, then $T_{rs} \approx 15$ nsec. If, on the other hand, the displayed rise time were 13 nsec, then the signal rise time would be $\approx$8.3 nsec. Measurements of rise times slower than about 5 times that of the oscilloscope do not have appreciable error if the instrument error is disregarded; signal rise times much faster than the oscilloscope rise time are subject to significant errors even if adjusted by Eq. (11-5-3) [5, 21, 22].

Low-frequency and High-frequency Oscilloscopes. General-purpose amplifiers are often divided into two classes: *low frequency* and *high frequency*. The low-frequency oscilloscope typically has video-frequency capability ranges (dc to $\approx$10 MHz), while high-frequency oscilloscopes often display single-shot signals with rise times as fast as 1 nsec ($\approx$350 MHz). Although the basic amplifier and time-base circuits are similar, low-frequency oscilloscopes often feature very low deflection factors (high sensitivity), while high-frequency oscilloscopes have good rise-time performance. Thus, low-frequency design specifications deal with noise levels, drift, CMR, and small nonlinearities since the signals to be measured are quite small. High-frequency design parameters include CRT writing speed, fast rise time, good high-frequency pulse response, and fast trigger capability.

High-frequency oscilloscopes are characterized by one or more of the following features seldom found in low-frequency oscilloscopes. Most frequently, a *delay line* will be in the high-frequency vertical unit to enable the operator to view the leading edge of a waveform (internal triggering). This is necessary for signals with rise times faster than about 50 nsec

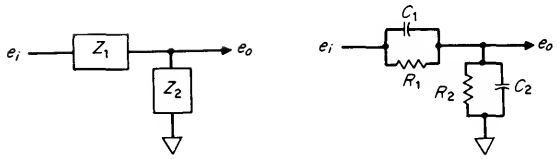
because of the time delay difference between the vertical amplifier and the time base. A postaccelerator CRT is usually required both to view high-speed waveforms with low repetition rates and to present adequately bright multichannel displays. Because of the added expense of post-accelerator CRTs and delay lines, the high-frequency oscilloscope is often partitioned into a display section, called a *mainframe*, with a plug-in compartment to accept a variety of relatively inexpensive *plug-ins* to give flexibility.

Amplifiers. Amplifiers are used on all three (x, y, z) deflection axes of the general-purpose oscilloscope.

The purpose of an amplifier is to provide gain between the input signal and the CRT so that small signals can be viewed and interpreted. Thus controls are necessary for calibrating the gain of the amplifier and also for modifying the gain to view signals of different amplitude. Additionally, the input impedance should be high enough so that the circuit or signal under test is not seriously loaded by the oscilloscope input impedance. Finally, because ac signals may be at quite arbitrary dc voltage levels, some controls for dc positioning or ac coupling are required.

A typical vertical amplifier in a laboratory oscilloscope may have a voltage gain of 2,000:1, preceded by an attenuator with ranges up to 500:1 or more. For example, if the CRT requires 20 V per division for deflection, by switching the attenuator the user may obtain an input deflection factor ranging from 10 mV to 5 V per division, usually in a 1, 2, 5, 10 sequence. A GAIN CAL control will permit amplifier gain adjustment (an accurate front-panel CALIBRATOR waveform is provided on most oscilloscopes to facilitate calibration). A VERNIER control allows continuous adjustment of the amplifier gain between calibrated attenuator settings. Once a given range is calibrated, attenuator accuracies will typically be within ± 3 percent on any other range.

Attenuators are designed to change the magnitude of the input signal seen at the amplifier input while presenting a constant impedance on all ranges at the attenuator input (which is the loading on the signal source). A compensated RC attenuator is required in order to attenuate all frequencies equally (Fig. 11-16). Without this compensation, high-frequency signal measurements would always have to take the input-circuit RC time constant into account. This would be very restrictive since the time constant would change on each range, and errors would be noticeable at frequencies as low as 3 kHz on many oscilloscopes. The input impedance for most oscilloscopes is 1 MΩ ± 1 percent, shunted by a small capacitance (typically between 10 and 80 pF), which leads to the term *high-impedance attenuators*. Very high frequency oscilloscopes sometimes use attenuators terminated by 50 Ω, which are more desirable at frequencies where transmission lines are used. Adjustments are usually provided on



$$Z_1(s) = \frac{R_1}{1 + sR_1C_1} \qquad Z_2(s) = \frac{R_2}{1 + sR_2C_2}$$

$$\frac{e_o}{e_i} = \frac{Z_2}{Z_1 + Z_2} = \frac{R_2(1 + sR_1C_1)}{R_2(1 + sR_1C_1) + R_1(1 + sR_2C_2)}$$

Choose $R_1C_1 = R_2C_2$, which gives

$$\frac{e_o}{e_i} = \frac{R_2}{R_1 + R_2}$$

with no frequency parameters

FIG 11-16 A single-stage compensated RC attenuator.

attenuators for both compensation and input capacitance on each range. One other point of interest on oscilloscope attenuators is their ability to be cascaded since each section has a 1 MΩ input (Fig. 11-17).

A CRT with electrostatic deflection has two vertical plates that are driven in a push-pull fashion by the amplifier. The amplifier may be designed as a differential amplifier from input to output, or it may be single ended at the input, converting to differential before reaching the

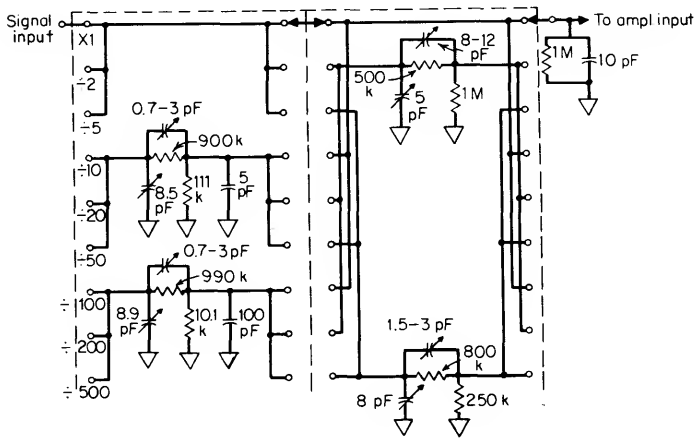


FIG 11-17 A typical high-impedance attenuator with nine cascaded ranges (providing division by 1, 2, 5, 10, 20, 50, 100, 200, and 500). Input impedance is 1 MΩ shunted by 10 pF on all ranges.

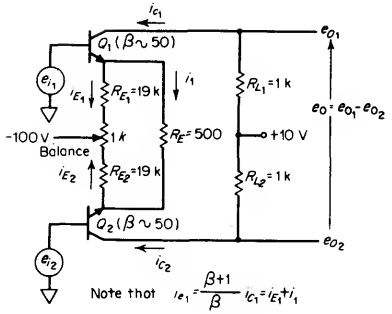


FIG 11-18 Differential amplifier.

CRT. Differential amplifiers are particularly valuable for oscilloscope amplifiers because of their capability in rejecting spurious signals, whether due to coupling through air from nearby sources or due to conduction. In addition, drift and stability characteristics are improved over a wide environmental operating range with differential amplifiers [23].

Refer to Fig. 11-18 for the operation of a differential amplifier. Assume first that e_{i1} and e_{i2} are zero. If the two sides of the amplifier are properly balanced, the output voltage $e_o = e_{o1} - e_{o2}$ will also be zero. Perfect balance requires that the V_{be} and β of each transistor be equal and that $R_{E1} = R_{E2}$, while $R_{L1} = R_{L2}$. Since such perfect combinations are improbable, a BALANCE control is inserted, as shown, for balance compensation. Note that i_1, through resistor R_E, is approximately zero for a balanced amplifier with no input signal. If e_{i1} is changed to $+1$ V, i_{E1} changes by about 1 percent with the very large current source, but the voltage across R_E is now 1 V, which gives 2 mA for i_1. The sum of $i_1 + i_{E1}$ must come from collector current for Q_1, which thus results in a signal change of -2 V on R_{L1}. Note that the gain at this point is twice the ratio R_{L1}/R_E. Also, through Q_2, the collector current now needs only to supply $i_{E2} - i_1$, which results in a $+2$ V change at e_{o2}. Here again the gain is ≈ 2, which results in a net gain of about 4, as one would expect for the Thevenin equivalent single-ended amplifier circuit.

Let both e_{i1} and e_{i2} be $+1$ V signals, and it is seen that i_{E1} and i_{E2} both change about 1 percent, but R_E still has no voltage across it since both sides have gone up 1 V. Thus i_{C1} and i_{C2} both change in the same direction about 1 percent, and the differential voltage output is still zero. If the circuit is not precisely balanced in all respects, some conversion of the common-mode signal may occur. For example, if β_{Q1} varies slightly (and β_{Q2} is constant), a small difference signal will be generated between e_{o1} and e_{o2}. Such conversion of the common-mode signal to differential signal is, of course, undesirable. The CMR ratio is a measure of the ability of a differential amplifier to avoid this conversion.

If a signal of $+X$ V is applied to both inputs of a differential amplifier, and some conversion to differential signal occurs, Y volts of differential output signal will result ($X > Y$). By contrast, if $+X$ V is put into one side of the amplifier and $-X$ V into the other side, the differential gain gives Z V of differential output ($Z > X$).

$$\text{CMR} = \frac{Y}{Z} = \frac{\text{differential conversion gain}}{\text{differential signal gain}} \tag{11-5-4}$$

Common-mode rejection is usually given in decibels, a -100-dB level indicating that a differential signal of 100 μV is displayed just as well as a common-mode signal of 10 V. Such ratios are especially important when measurements of low-level signals (such as strain gage or medical transducer signals) are made in the presence of large electromagnetic fields (such as 60-Hz power transformers). Common-mode rejection is usually degraded by inclusion of attenuators and other front-panel controls such as ac coupling, as well as at higher frequencies [24].

One fact affecting both amplifier gain and balance is that transistor parameters such as β and V_{be} are sensitive to temperature. The differential amplifier helps to minimize this problem since a change in β or V_{be} caused by temperature variation will likely affect both devices and the change in gain or balance will thus be primarily common-mode change. Only differential changes will affect the output, and these are very small effects by comparison.

Differential amplifiers are of considerable value for accurate comparator measurements. In such tests, a standard signal may be fed into the negative vertical input and the unknown signal into the positive input. The on-screen presentation then represents the difference between the two signals. Much greater accuracy may be obtained in this manner than by observing the entire unknown signal on-screen, where resolution is usually about 1 or 2 percent. Circuits designed expressly for comparator measurements offer very high dynamic range (perhaps 10,000 divisions of effective input signal without distortion of the compared signals), low CMR ratio (< -80 dB from dc to perhaps 50 kHz or more), and an accurate comparator reference voltage, which may be used to null the unknown input signal. Such units are of necessity fully differential amplifiers, with very closely balanced parameters. Measurements with such amplifiers may approach 0.1 percent absolute accuracy, with relative accuracy limited primarily by the reference standard [25].

Multiple-trace displays are of great advantage in oscilloscopes, for they enable the operator to make time and amplitude comparisons among several waveforms. Such measurements, useful in production test and service areas, are indispensable for many laboratory tests. The operator usually desires the horizontal time scale to be the same for all channels

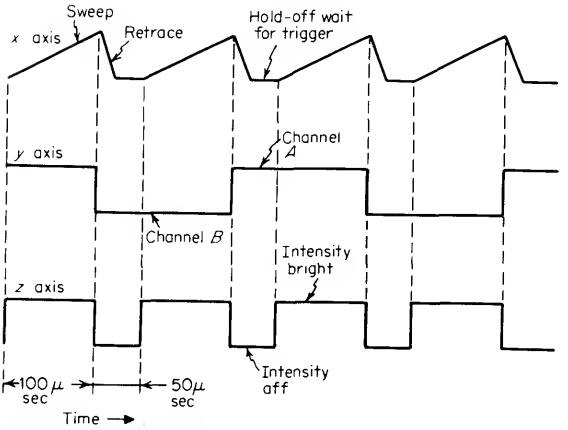


FIG 11-19 Timing relations of a dual-channel vertical amplifier in ALTERNATE operation.

while the attenuators and position controls of the vertical channels are independently controlled. Both dual-trace and four-trace vertical displays are commonly available.

Multichannel displays are commonly generated in two time-multiplex modes. The ALTERNATE mode displays one vertical channel for a full sweep, and the next vertical channel on the next sweep. Thus the vertical channels are alternately displayed (Fig. 11-19). The CHOP mode allows a small time segment of a given sweep to be allotted to the first vertical channel, and the next time segment is allotted to the second vertical channel. Thus, the vertical channels are composed of small "chopped" segments, which merge to appear continuous to the eye (Fig. 11-20) if enough chopped segments are included.

The choice of CHOP or ALTERNATE for multichannel vertical displays is usually available to the scope operator. The ALTERNATE mode is commonly used for high-frequency signals, where sweep speeds are much faster than CRT phosphor-decay characteristics. CHOP mode is more useful at low sweep rates where the flicker effects of ALTERNATE mode are noticeable and objectionable. Two or three overlapping ranges where either mode is equally satisfactory are commonly provided.

The easiest way to obtain a dual-channel vertical system is to provide an electronic switch between two vertical preamplifiers (containing attenuators and positioning) and one vertical deflection amplifier (Fig. 11-21). This allows for time sharing the beam between the two preamplifiers as desired. Several modes of operation may be derived from such a dual-channel amplifier: CHANNEL A only, CHANNEL B only, DUAL-CHANNEL

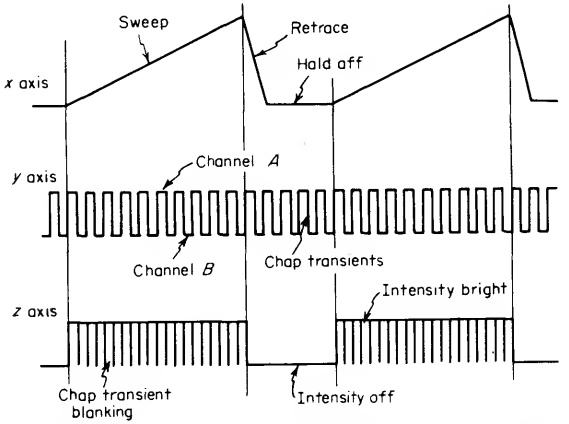


FIG 11-20 Timing relations of a dual-channel vertical amplifier in CHOP operation.

(either ALTERNATE or CHOP), the sum of $A + B$, or the difference $A - B$.

In dual-channel displays, the $A - B$ differential mode serves as a single-channel differential amplifier. For computer logic testing, an $A + B$ summing mode is frequently necessary. Two types of dual-channel switching arrangements are commonly encountered, each with compromises (Fig. 11-22). The first type, a true differential amplifier, obtains the $A - B$ differential mode by simply switching the channel B input and attenuator to the negative side of the channel A differential

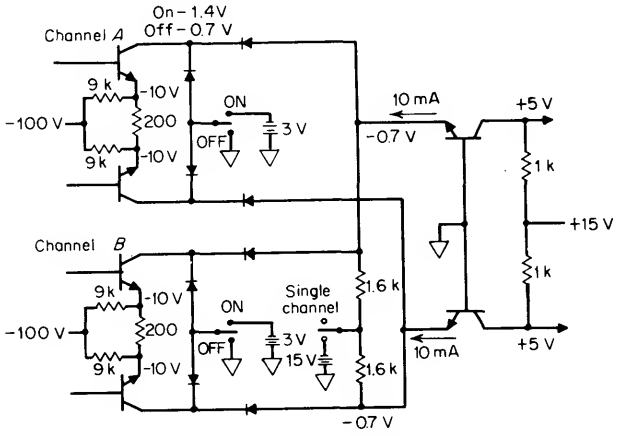
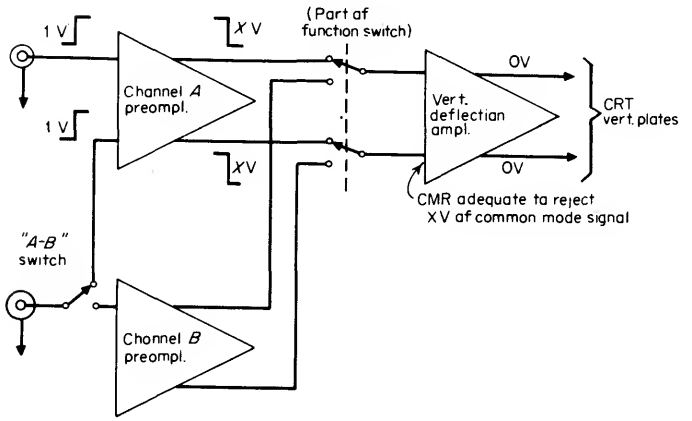
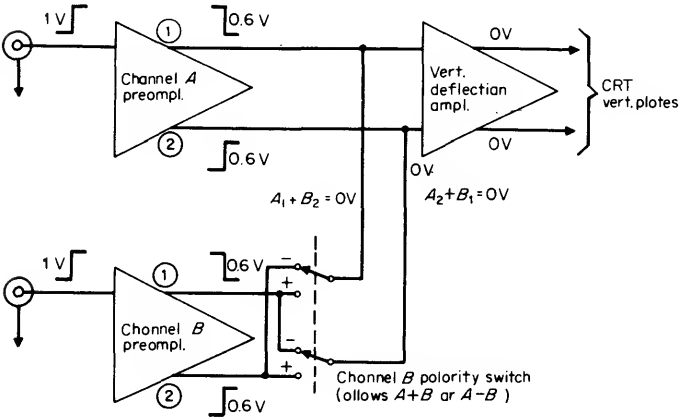


FIG 11-21 A dual-channel electronic switch.



(a) A−B mode, differentiol omp.



(b) A−B mode, summing ompl.

FIG 11-22 Dual-channel amplifier switching to accomplish differential ($A - B$) amplifier operation.

amplifier; this permits virtually all the CMR capability of a similar single-channel differential amplifier, but it does not allow an additive $A + B$ mode. The second type, a *summing* amplifier, uses either additive or subtractive logic at the same switching mode that is used for dual-channel switching and thereby obtains either $A + B$ or $A - B$ modes. One limitation here is the often substantial increase of CMR ratio due to unbalances occurring in any one of four sides of the two differential preamplifiers. Commonly, only −40 dB of CMR is available in the

$A - B$ mode for units of the second kind, whereas -60 dB is achievable for comparable effort with the first type. Also, the first type typically is able to achieve a wider dynamic range for overdrive common-mode signals at the input.

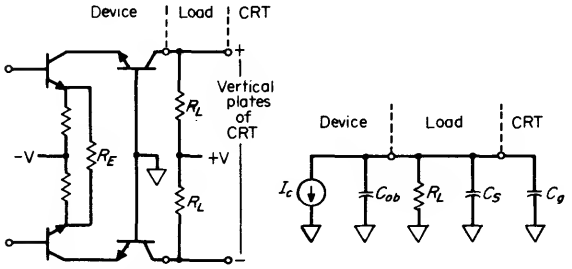
Deflection Amplifiers. Deflection amplifiers for electrostatic deflection systems have a number of restrictive design parameters in high-frequency oscilloscopes. The gain-bandwidth product of the output device, f_T, is often a determining factor in the ultimate deflection speed. Usually the constraints imposed by the CRT capacitance C_p and the deflection factor V_D V per division are even more important in determining deflection speed. For very high speed systems, distributed-plate CRTs are made in order to replace the capacitance by a transmission line of constant impedance.

The speed limits of a nondistributed-plate high-frequency CRT and amplifier can be discussed by using the single-ended equivalent circuit of Fig. 11-23.

An effective capacitance to ground C_g exists, on each CRT deflection pin, which must be charged or discharged in order to deflect the beam. The deflection amplifier device will also have an output capacitance C_{ob} and some stray capacitance from leads C_s may be included as well. Since a voltage change on a capacitor is governed by the time integral of current

$$v_c = \frac{1}{C_T} \int i_c \, dt \tag{11-5-5}$$

we see that the larger the total capacitance $C_g + C_{ob} + C_s = C_T$, the more current is required to obtain the same voltage deflection in the same time interval. Since center screen v_c times i_c equals the power dissipation P_D of the device, an equation for the minimum full-scale deflection time may



(a) Differential deflection amplifier

(b) Equivalent single-ended circuit

FIG 11-23 Deflection-amplifier frequency parameters.

be written as a function of C_T, V_D, and P_D. If the total deflection-amplifier dynamic range is k_1 divisions, the maximum deflection of one side is $k_1/2$ divisions, which gives a maximum power dissipation of

$$P_D \approx \frac{k_1 V_D i_c}{2} \tag{11-5-6}$$

If we assume a k_2-division vertical-deflection scan over which rise time is measured, and a peaked current supply i_c to maintain a linear rise, the amplifier output will have an approximate rise time of

$$T_r \approx \frac{k_2 V_D C_T}{2 i_c} \tag{11-5-7}$$

Thus, the rise time for the deflection system may be expressed

$$T_r \approx \frac{k_1 k_2 V_D{}^2 (C_g + C_{ob} + C_s)}{4 P_D} \tag{11-5-8}$$

Controls are difficult to design, for there are many intrinsic problems. The sequence of controls within an amplifier greatly affects their interaction and their consequent use for modifying signal displays on screen. For example, if the POSITION control precedes the GAIN VERNIER, any vernier rotation attenuates a dc position offset toward center screen. If the controls are reversed, no position shift occurs relative to center screen when gain is changed. Instead, signals are attenuated with respect to signal ground. The difficulty for the designer comes about because each sequence is of value for certain kinds of display (Fig. 11-24), and since either method may be found on contemporary scopes, the user is faced with two sets of identically labeled controls performing different functions on two otherwise similar scopes. Since analogous problems are encountered with the location of polarity switches, trigger signal derivations, and interstage attenuators, control function design is one of the greatest variations found in oscilloscopes today.

Time Bases. Time-domain oscilloscopes require a sweep generator that is linear with time for the x-axis display. Such sweep circuits combined with display gating functions are termed *time bases*. A high-quality time base may feature sweep-time variations from 10 nsec to 5 sec per division, with time accuracy from range to range of better than 3 percent and linearity better than to within 1 percent across the CRT, and a 10-times expansion in the horizontal amplifier to allow 1 nsec per division displays for very high speed transients. Other x-axis sweep generators of occasional value in special-purpose oscilloscopes include sinusoidal, parabolic, exponential, and hyperbolic generators.

A sawtooth waveform differentially applied to the horizontal deflection

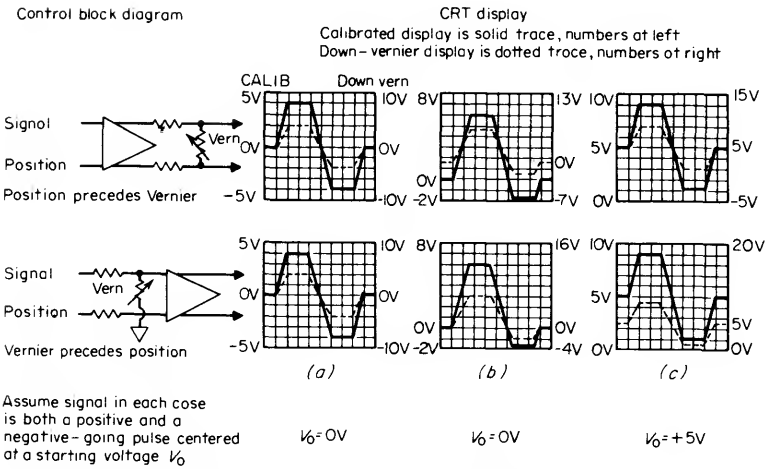


FIG 11-24 Interaction of POSITION and GAIN VERNIER controls. Note that both control arrangements are equivalent at left (*a*), while (*b*) and (*c*) vary. The lower arrangement is usually preferred for (*b*), while the upper is often favored for (*c*).

plates will move the beam back and forth across the CRT screen. As the beam goes from left to right, a linear time sweep is traced. A signal to the gate amplifier turns on the *z*-axis intensity, which allows the sweep trace to be displayed. The right-to-left movement, called *retrace*, or *flyback*, is blanked out. A constant-current source charging a capacitor will produce the desired sawtooth waveform if a discharge path is provided (Fig. 11-25). Relaxation oscillators, with neon tubes or unijunction transistors, provide an inexpensive sawtooth generator with low accuracy; bootstrap techniques allow much greater linearity with additional expense.

The *Miller integrator* is the most common time-base generator in laboratory oscilloscopes today (Fig. 11-26). It basically is an operational

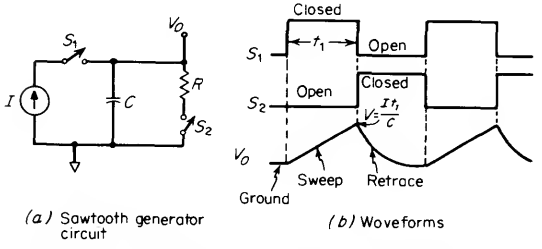


FIG 11-25 A simple ramp or linear sweep generator.

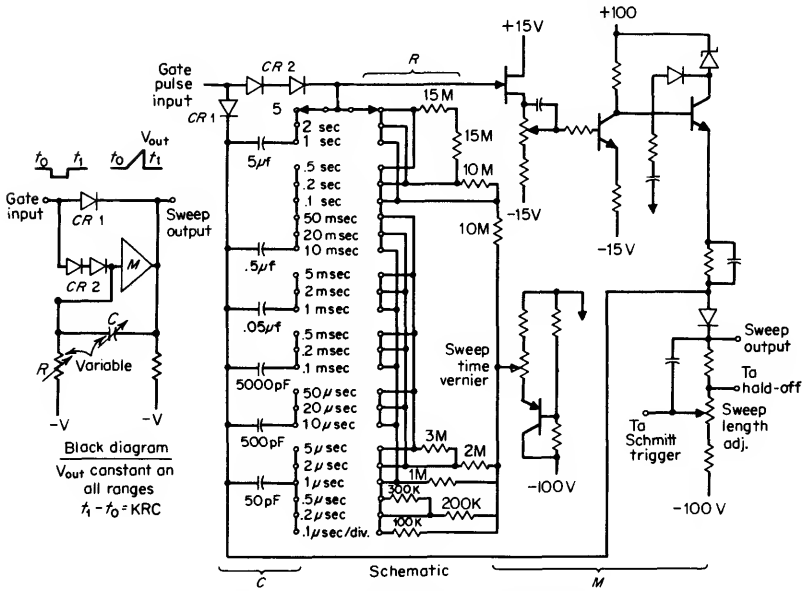


FIG 11-26 A Miller integrator with 24 selectable sweep speeds (100 nsec per division to 5 sec per division in a 1, 2, 5, 10 sequence).

amplifier with the Miller feedback capacitance to convert a step-function input into a sawtooth output. Historically, special-purpose vacuum tubes such as the phantastron, sanatron, and sanaphant were developed to facilitate time-base generator design based on the Miller integrator. Today, designs usually employ standard bipolar and field-effect transistors [27, 28].

One of the great advantages of the Miller integrator is its flexibility in choice of both R and C in the feedback loop. The ranges are such that a high-input-impedance device (such as a vacuum tube or field-effect transistor) will permit scaling C from 10 pF to 1 μF or more, and R from 100 kΩ to 50 MΩ, which allows sweep speeds to vary, for example, from 100 nsec to 5 sec per division with the same generator.

Additional components of the time base include a *trigger generator* to convert the signal from the vertical amplifier, or external source, into trigger pulses, which turn on the *gate generator* to start the Miller integrator, a *hold-off circuit* to allow recovery of the Miller integrator after a sweep is completed, and a *reset Schmitt-trigger* circuit which rearms the gate generator when the hold-off time is completed. A block diagram of a time base is shown in Fig. 11-27, and a waveform sequence of events in Fig. 11-28.

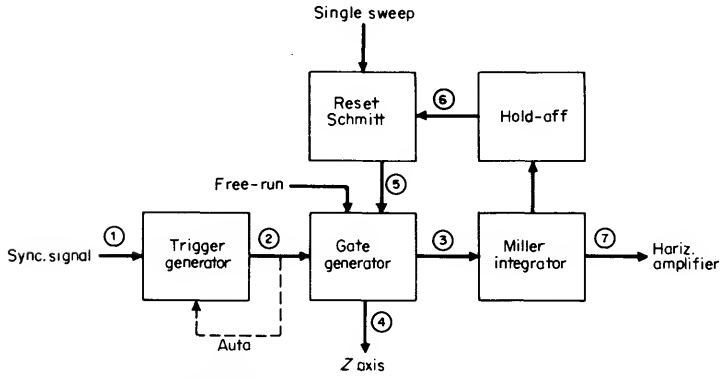


FIG 11-27 Time-base block diagram.

Other common sweep circuits include an AUTO circuit which senses the length of time after a sweep occurs and automatically provides a trigger pulse if no signal has come from either the vertical amplifier or an external source after approximately 20 msec, or a time determined by the SWEEP TIME setting. This function thus gives a base-line presentation without flicker in the absence of a signal and allows easy verification of ground level or a *base-line* of a voltage level on the CRT. A signal from the gate circuit is sent to the *z*-axis blanking amplifier to turn the intensity up when a sweep cycle starts, and off when the sweep recovery begins. A FREE-RUN mode is possible, which restarts the Miller integrator immediately after the hold-off time. A SINGLE-SWEEP mode is possible also, which allows the sweep to run once from a trigger pulse after which the gate generator is not reset until the operator desires. This mode is of

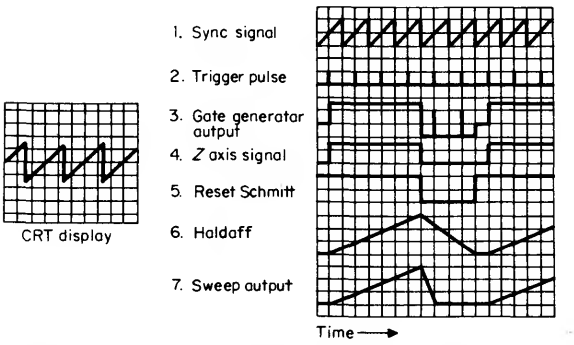


FIG 11-28 Time-base waveforms (waveform numbers correlate to points on Fig. 11-27).

prime value for displaying and photographing single-occurrence events when the time of occurrence is undetermined.

Sweep Modes. A second time-base generator is supplied for some oscilloscopes. If the CRT is a dual-gun tube (two independent writing beams), the sweeps are truly independent. The same is true for two vertical channels. Multigun tubes are of chief value in an application requiring writing of nonsynchronous information simultaneously (as heart-rate monitoring of two patients) or display of two different single-shot phenomena occurring simultaneously (such as two points of impact in a collision).

A broad variety of measurements formerly accomplished by dual-beam oscilloscopes is being done today by single-beam oscilloscopes with a multiplexed pair of time-base generators. Depending upon the sophistication of the second time-base generator, one or more of three additional

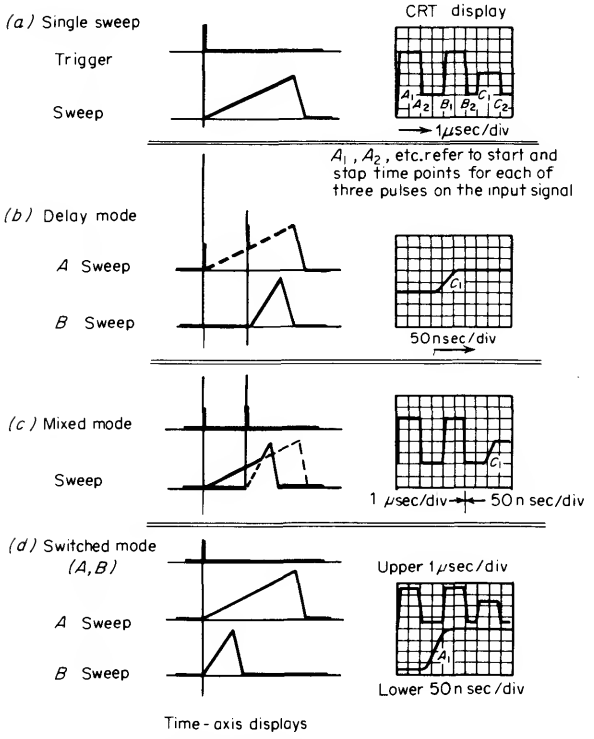


FIG 11-29 Timing relations and CRT displays of four common sweep modes.

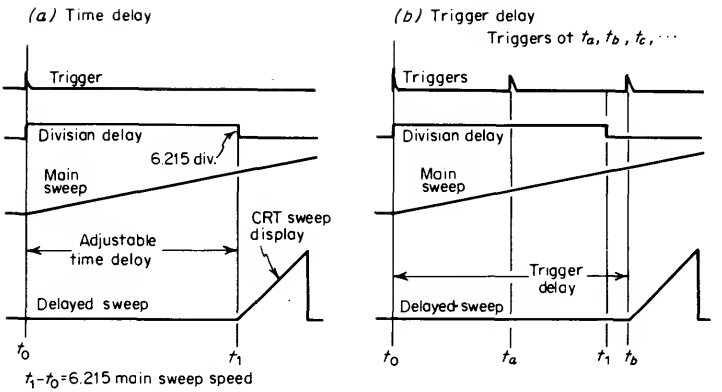


FIG 11-30 DELAY mode operation showing the time sequence of MAIN (delaying) SWEEP and DELAYED SWEEP.

time-axis displays may be obtained. These are called DELAYED SWEEP [29], MIXED SWEEP, or SWITCHED SWEEP (Fig. 11-29).

The simplest and most common of the three is DELAYED SWEEP. Two time-base generators are incorporated, but the display is only of the second time-base generator. The first time-base generator is called a *delaying* generator and the second is termed a *delayed* generator. The time-base ramp generated by the delaying generator is started by a trigger pulse at time t_0; it continues until reaching a comparator level set by the DIVISION DELAY control. At this time t_1, the delaying generator stops, and the delayed generator is ready for operation. Two modes are now possible, depending upon the trigger setting for the delayed generator. If the delayed generator is in AUTO, it runs as soon as the delaying sweep reaches the DIVISION-DELAY comparator level, as shown in Fig. 11-30. This mode uses the first generator as a *time-delay generator*. If it is being triggered from an internal or external trigger pulse occurring at time t_b (after time t_1), then the delaying generator is being used as a *trigger-delay generator*. This trigger-delay mode is often called *arming*, for it serves to ready the delay generator to run at the next trigger pulse. This is of great significance for a common pulse-train measurement situation where the pulses are subject to time jitter between t_0 and t_b.

DELAYED SWEEP offers increased accuracy and resolution for many time-interval measurements. The DIVISION DELAY control [termed variously DELAY-TIME MULTIPLIER, DELAY (DIV), or DELAY (cm)] is the key control for such measurements. Normally, a 10-turn potentiometer with a readout of three significant numbers and a vernier scale is provided for the knob control.

Measurements may be made in several ways, with varying accuracy capability. If a single pulse is to be measured for time delay from a reference t_0, the DELAY knob may be dialed until the pulse is centered on screen, and the pulse time delay is then calculated. If, for example, the main SWEEP TIME is 10 μsec per division, and the DIVISION DELAY setting is 6.215 divisions, the pulse time delay is 62.15 μsec (± 3 percent for sweep-time accuracy) or 62.2 μsec ± 1.9 μsec.

If a reference time-mark pulse is used in conjunction with the pulse to be measured (with a dual-channel vertical scope), the accuracy may be increased substantially. Assume, for example, that a reference pulse occurs at $t_0 + 50$ μsec. If the delaying SWEEP TIME is set to 2 μsec per division, the CRT display should show the two pulses separated by 6.1 divisions, which then gives a difference reading of 12.2 μsec ± 3 percent for sweep-time accuracy and ± 3 percent for deflection nonlinearities and resolution. This results in an overall measurement of 62.2 sec $\pm$ 0.7 μsec. Further improvement could result if the DIVISION DELAY knob is used first to center the reference pulse, and then to center the pulse of interest, for the deflection nonlinearity is canceled. This gives a measurement of 62.2 μsec $\pm$ 0.4 μsec. Time differences between any two pulses can be handled in a similar fashion [29].

The MIXED mode provides a more convenient display pattern for many time-interval measurements, since both the reference signal and the signals of interest may be viewed in a wavetrain context with only the region of interest expanded for time-interval measurement. The MIXED mode is very similar functionally to the DELAY mode, except that the delaying sweep generator is displayed first. An external trigger is generated at time t_1, whereupon the delayed sweep generator (at a faster sweep speed) is started. Once it reaches the same ramp height as the delaying generator, it is displayed instead for the rest of the sweep cycle. This mode is especially valuable in pulse-train studies, when the time interval of a specific set of pulses in a complex pulse-code train is of interest.

DUAL SWEEP or SWITCHED SWEEP mode displays two independently variable sweep speeds by electronically time sharing between the two time bases (Fig. 11-31). Depending upon sweep speeds, the electronic switching is accomplished similarly to the CHOP and ALTERNATE modes discussed earlier under vertical amplifiers. Some difficulty is encountered in practice with beam brightness (variations and flicker) if the two channels vary by much more than an order of magnitude in sweep times.

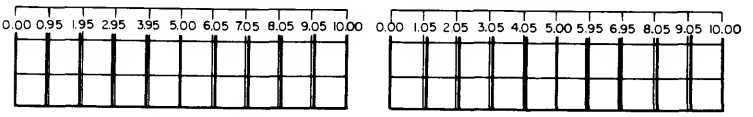
Sweep Linearity. The various scope manufacturers have failed to define sweep linearity consistently, and this has resulted in some confusion regarding sweep linearity specifications. Figure 11-32 illustrates two different 10-division sweep displays with 10 equal time intervals from

FIG 11-31 A switched-sweep scope displaying three vertical channels simultaneously on two different sweep speeds. (*Tektronix, Inc.*)

a time-mark generator. The calculations shown are derivations of three different sweep linearity specifications in common use.

The first method defines the sweep linearity error to be the maximum deviation of any time mark from a graticule mark divided by the time-mark interval. The second method, which gives a larger error number (and perhaps is in disfavor for this alone), involves measuring the maximum and minimum interval widths, then dividing the difference by the average interval width. The third method averages the difference between maximum and minimum interval widths. Inconsistent sweep nonlinearities as shown in Fig. 11-32*b* are not reflected by a specification based upon the first method; this is probably its major limitation. If a sweep linearity error is of the form shown in Fig. 11-32*a*, which is a more typical form, then any of the three methods give comparable results. Note also that each method includes linearity errors in the horizontal deflection amplifier and CRT as well as the time base.

Trigger Generators. The circuit block responsible for starting the sweep at the desired point on a waveform is called the *trigger generator*. Trigger generators incorporate a selection of several trigger sources, plus a variable



Linearity measurement

Maximum deviation between a time mark and a graticule mark = 0.05 div, average time-mark interval = 1.00 div, linearity error = $\frac{0.05}{1.00} = 5\%$	Method 1: Maximum deviation of any time mark from a graticule mark divided by the average time-mark interval	Maximum deviation between a time mark and a graticule mark = 0.05 div, average time-mark interval = 1.00 div, linearity error = $\frac{0.05}{1.00} = 5\%$
Smallest time-mark interval = 0.95, largest time-mark interval = 1.05, average time-mark interval = 1.00, linearity error = $\frac{1.05 - 0.95}{1.00} = 10\%$	Method 2: Maximum difference between smallest time-mark interval and largest time-mark interval divided by the average time-mark interval	Smallest time-mark interval = 0.95, largest time-mark interval = 1.10, average time-mark interval = 100, linearity error = $\frac{1.10 - 0.95}{1.00} = 15\%$
Total time-mark interval = 10.00, linearity error = $\frac{1.05 - 0.95}{10.00} =$ 1%	Method 3: Same as method 2, except divided by the total time-mark interval	Total time-mark interval = 10.00, linearity error = $\frac{1.10 - 0.95}{10.00} =$ 1.5%

FIG 11-32 Sweep linearity.

comparator to set the desired trigger level and a trigger pulse generator which starts the sweep generator (Fig. 11-33).

The three typical trigger sources are INTERNAL, EXTERNAL, and LINE. The internal trigger source provides a replica of the signal applied to the vertical amplifier; the external trigger source is derived from an external input; and the line source is derived from the frequency of the power line. With these three sources, the controls for ac or dc coupling, and the polarity selection, one should be able to derive a signal that will give a stable display on the screen. An AC FAST mode is sometimes provided to give a high-frequency trigger capability in the presence of low-frequency signals; AC SLOW gives low-frequency triggering in the presence of high-frequency signals.

The trigger level control adjusts the comparator threshold level to allow selection of the input voltage level (of either polarity) which will switch the differential comparator. This produces a phase-controlled or

time-controlled trigger pulse at the output. Since only one trigger pulse is provided for each waveform cycle, and it is always at the same phase angle, the sweep will be triggered each time at the same relative phase angle. Consequently, the repetitive beam traces on the screen will all start at the same point, and the trace will not jitter or appear unstable. Jitter at high frequency sometimes occurs because of the finite rise time of the trigger pulse plus the finite reset time of the time-base gate generator. Tunnel diode pulse circuits in both areas have minimized this problem in recent designs because of their very fast switching characteristics [30].

Complex pulse train signals, such as those found in computer logic and pulse-code-modulation communications, are very difficult to display on a scope because of the difficulty of developing a satisfactory trigger signal. Figure 11-34 illustrates a complex pulse train of seven pulses for which a stable scope display is desired. For the repetition rate shown, a standard SWEEP TIME of neither 5 nor 10 μsec per division will allow a synchronous trigger signal to be developed. Thus the displays will indicate a number of places of double pulses as shown. One of the three ways illustrated may be used to synchronize the display. The SWEEP VERNIER control will give a stable display whenever the total sweep and hold-off time approaches the fundamental repetition rate or a subharmonic of it. The display is uncalibrated in time, however. The SWEEP LENGTH control works in a similar manner, seeking a time correlation between the repetition rate and the SWEEP TIME with hold-off time. The display remains calibrated in time, but the total sweep length becomes shorter and gives a smaller display window. A VARIABLE HOLD-OFF control will

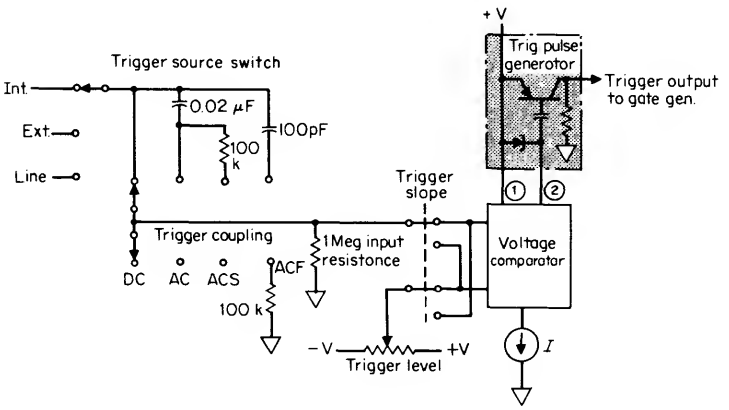


FIG 11-33 Trigger-generator circuit block.

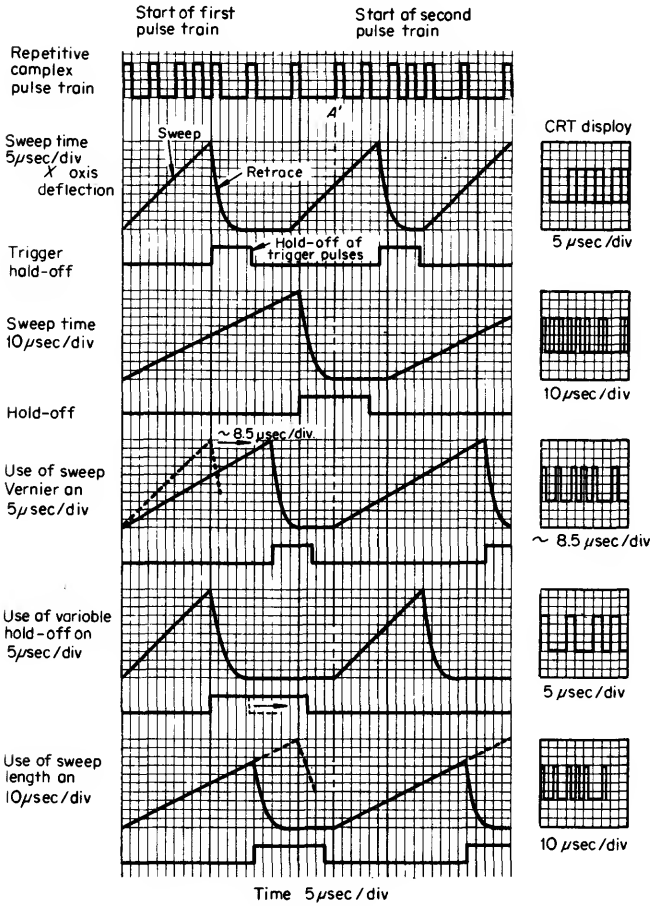


FIG 11-34 Complex pulse-train triggering.

allow the calibrated full horizontal display to be retained while a synchronous trigger is obtained. If the range of the hold-off variation is greater than the total sweep time (including recovery time), stable triggering is ensured [31].

Delay Lines. High-frequency oscilloscopes nearly always include delay lines in the vertical amplifiers. The purpose of such lines is to delay the vertical signal enough to keep it from reaching the CRT deflection plates before the horizontal sweep circuits are running. As shown in Fig. 11-35, the vertical signal triggers the sweep generator and enables the horizontal amplifier to begin tracing a sweep at full intensity before the vertical

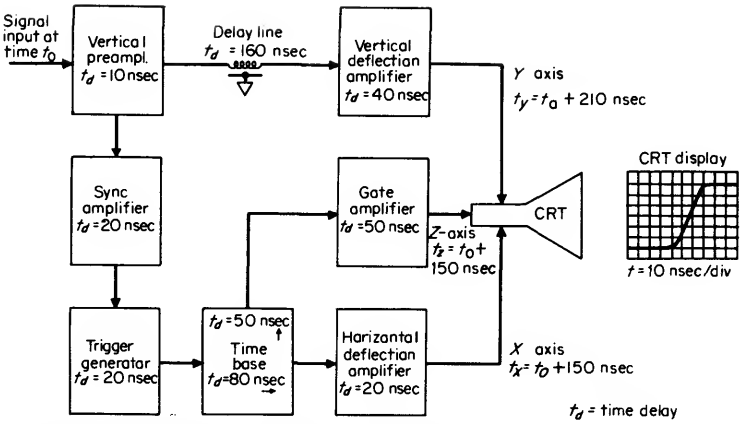


FIG 11-35 Delay-time relationships in a high-frequency oscilloscope block diagram.

signal reaches the vertical deflection plates of the CRT. Thus, the first part of the signal is displayed, and even if it is a single-shot phenomenon, it may be viewed much like any other signal.

The first delay lines in oscilloscopes were most frequently lumped-parameter lines, as in Fig. 11-36*a*, sometimes numbering 50 cascaded

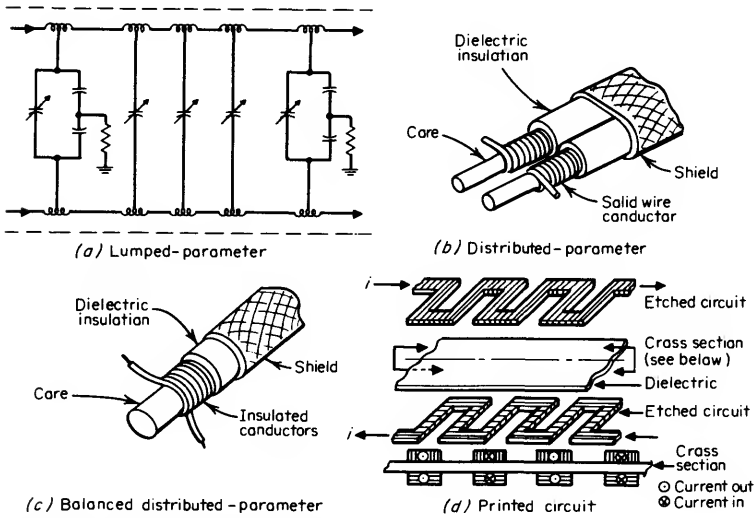


FIG 11-36 Delay-line construction techniques.

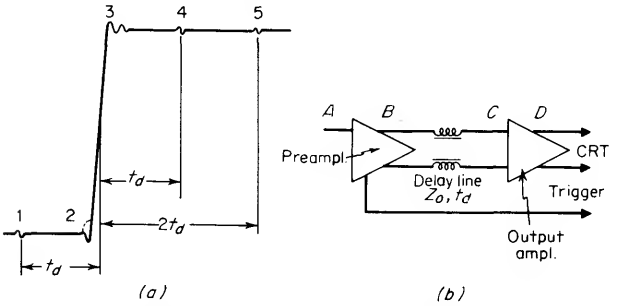


FIG 11-37 Delay-line pulse-response perturbations: (*a*) pulse response, (*b*) block diagram.

LC delay segments. These were iterative T sections, each section capable of delaying the signal between 3 and 6 nsec. Since the amplifier was often differential at this point to minimize trace drift, two segmented lines had to be adjusted and matched in both pulse response and delay time. Servicing these lines with up to 68 adjustments was difficult at best, but they were capable of precise performance when adjusted properly [32].

A distributed-parameter line (essentially a transmission line), constructed with a wound helical coil on a mandrel and extruded insulation between it and an outside shield, was developed as an alternative approach. Two of these lines are used for differential circuits. A minor refinement puts both cores, wound in opposite directions, inside one shield, Fig. 11-36*b*, which helps to equalize the distributed capacitance between the two lines. A later improvement has a winding of two insulated wires in a shoestring interlace on the same mandrel; then an insulation and shield is placed over both wires, as in Fig. 11-36*c*. This results in the most uniform balance between the two lines accomplished to date.

Etched-copper delay lines on circuit boards have also been developed in recent years. See Fig. 11-36*d*. Besides good uniformity, advantages of this approach include low cost of materials, high yield, and greater reliability [32, 33].

Implementation of delay lines in oscilloscopes requires careful attention to impedance matching to avoid reflections. Radiation of signals around the line will also lead to undesirable display anomalies. Five common problems occurring in amplifier pulse responses that are due to the delay line are shown in Fig. 11-37. Anomaly 2, called *preshoot*, is a discontinuity caused by improper phase delay, as may be the *ringing overshoot* of anomaly 3. These are both functions of the delay-line design or the terminations. Figure 11-37*b* illustrates the vertical-amplifier block diagram, which aids in explaining the other distortions. The first

anomaly results from an undelayed signal's coupling around the delay line, which avoids the delay time t_d. This occurs either by capacitive coupling or radiation from points B or C, or possibly by coupling from the trigger-generator lead into any part of the deflection amplifier. The fourth anomaly, one delay time later than the rise-time delay, is due to radiation or coupling from the deflection amplifier back into the preamplifier. The large CRT deflection signal at point D is often coupled through power supply lines back into point A to cause this distortion. The final disturbance is a mismatched-delay-line termination at both ends of the line. If a signal travels through the line to point C and meets a discontinuity, it reflects back through the line (in $1t_d$). If at the other end, point B, it encounters a second mismatch, it travels again to point C, arriving with the distortion seen at $2t_d$. A perfect termination at either end would stop this distortion, but in practice these terminations are very difficult to achieve. All the distortions discussed are typically held to less than 1 percent of the signal amplitude, and usually they are very difficult to discern.

Plug-in Oscilloscopes. The plug-in feature of many oscilloscopes allows great flexibility in changing the measurement capability of the instrument. Various interface levels for plug-ins have been developed that determine both the flexibility and efficiency of the main-frame oscilloscope. Units having the interface at the CRT are the most flexible in display capability; main frames in this class will accept plug-ins which give a variety of low and high sensitivity, single and multichannel and low- and high-speed vertical and horizontal functions not only for oscilloscopes but also for spectrum analyzers, operational amplifiers, medical instruments such as vector cardiology plotters and myographs, and time-domain reflectometry units. Units having the time-base and vertical delay line and deflection amplifiers incorporated in the main frame are versatile only in vertical preamplifier capability. On the other hand, economies in instrument-system cost are realized, which may better serve a customer who is only interested in traditional oscilloscope measurements. This occurs, for example, because the customer buying three high-frequency vertical plug-ins (such as a single-channel low CMR unit, a dual-channel wideband unit, and a four-channel logic unit) does not have to buy three delay lines.

A compromise approach (Fig. 11-38) includes the horizontal amplifier and gate amplifier in the main frame (to avoid duplication in each plug-in), but a full vertical plug-in and a time-base plug-in option to allow flexibility in frequency response and other amplifier characteristics. The absence of a delay line in the main frame and the inclusion of external horizontal and gate amplifier inputs allow easy use of the main frame for xy plotting with video information.

FIG 11-38 A high-frequency oscilloscope with plug-ins for both x and y axes. (*Hewlett-Packard Company.*)

Portable Oscilloscopes. Because of the complexity of oscilloscope circuits and the size of CRTs, the oscilloscope has traditionally been a large instrument with limited portability. In recent years, solid-state designs have permitted size reduction of general-purpose oscilloscopes. The plug-in scope of Fig. 11-38, for example, weighs about 30 lb with plug-ins, and yet it has better performance than many recent scopes with greater than twice the weight and volume. The need for a light-weight field-service scope has led also to a special breed of non-plug-in high-performance portable scopes, which are capable of bandwidths and sweep-time capabilities fully equivalent to any general-purpose equipment (Fig. 11-39). Low-power circuitry allows some portable scopes to be operated from batteries rather than 110-V ac power. Rugged constructions and very small size make these instruments attractive in military or industrial field situations. Dual-channel vertical amplifiers with medium sensitivity and a delaying-sweep time base are typically provided in such units, and where the small viewing area and lack of flexibility are not handicaps, such units provide very good measurement capability.

FIG 11-39 A high-frequency field-service oscilloscope. (*Tektronix, Inc.*)

Storage Oscilloscopes. Three common kinds of display measurements, very difficult to view on a standard oscilloscope CRT, are readily displayed on a *storage* oscilloscope tube. Single-shot events (such as the waveform of an explosion) are transient in nature and quickly lost to the observer unless the waveform can be either photographed or stored on the CRT itself. Signals with very low repetition rates are also difficult to view on standard oscilloscopes because of phosphor decay. This is most annoying in medical displays, mechanical motions, and radar displays, three prominent areas of application for CRT indicators. Lastly, comparison of two waveforms is often difficult on a standard oscilloscope since CRT refresh characteristics do require simultaneous testing of the two circuits. In contrast, being able to store one signal trace permits comparison with a

FIG 11-40 A low-frequency variable-persistence oscilloscope. (*Hewlett-Packard Company.*)

second signal obtained at some later point in time. Production-line calibration, especially in iterative adjustments or testing for previously established waveform limits, often requires such a display capability. For each of these important applications, the storage oscilloscope amply justifies its additional cost and complexity for many users.

Storage oscilloscopes to date have been available primarily as main-frame displays accepting a general-purpose line of plug-ins. The plug-in lines in both Figs. 11-30 and 11-38 include storage main frames, for example. Thus the unique display properties of the storage CRT are mated with a wide variety of amplifiers, time bases, and even spectrum-analyzer plug-ins. Both phosphor-storage and mesh-storage tubes are available in storage-oscilloscope main frames, and recently some non-plug-in low-frequency mesh-storage oscilloscopes have been introduced that have variable persistence (Fig. 11-40).

11-6 Sampling Oscilloscopes

In the past decade, computer design, pulse-code telemetry, and the similar high-speed information systems have generated rapidly expanding requirements for observing signals with wider bandwidth, such as shorter, sharper pulses, and their timing correlation.

Recent advances in solid-state devices have made subnanosecond pulses possible, but until a few years ago, there was no convenient way to analyze them. Real-time scopes with traveling-wave tubes had been developed for such speeds, with the vertical signal coupled to the display tube either directly or through a distributed amplifier [22]. Scopes with 1 GHz or more y-axis bandwidth were produced by such techniques, and while they were unquestionably of value, they were seriously handicapped by small CRT display area, low sensitivity, and poor brightness on low-duty-cycle displays.

The sampling technique avoids the problems inherent in high-frequency real-time scopes by translating high-frequency signals to a lower-frequency domain. The sampling oscilloscope relies upon a technique very similar to that of a stroboscopic light in providing visual observation of rapid motion. Instead of continuously monitoring the signal under test, the sampling device of the oscilloscope "samples" the signal amplitude at regulated intervals, and synthetically reproduces the sampled signal. These samples are presented on the CRT as a series of dots (Fig. 11-41). Many thousands of dots may be displayed across the CRT screen, and they merge to appear as a continuous line to the observer. Both a stroboscope and a sampling scope require a repetitive waveform in order to portray an apparent image. Both also depend upon the capability of the

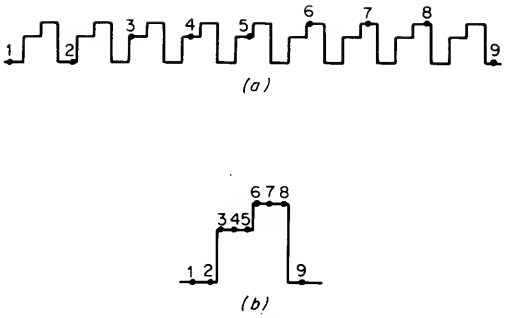


FIG 11-41 Reconstruction of a signal with sampled dots: (*a*) signal being sampled; (*b*) presentation on CRT.

human eye or a CRT phosphor to respond to very rapid pulses of information and to store the image between pulses so that it appears continuous.

The sampling technique allows the design of an oscilloscope with wide bandwidth, high sensitivity, and a bright, clear display even for relatively low-duty-cycle pulses. The sampling technique allows nearly all the flexibility of any real-time scope plus the virtue of much higher effective frequency capability whenever a repetitive signal is to be observed. Thus the very high frequency (>300-MHz) real-time scopes are today used primarily with single-occurrence phenomena.

Sampling as a measurement technique for high-speed electrical signals was used as early as 1849, far prior to the first cathode-ray oscilloscope. During intervening years, improvements were made which resulted in an instrument termed an *ondograph*, produced for the first three decades of this century. Since 1950, when J. M. L. Janssen built a 35-MHz *stroboscopic* oscilloscope to demonstrate applicability of the sampling concept to an oscilloscope display tube, there has been notable progress in reintroducing sampling techniques to high-speed electrical-signal analysis. Sampling oscilloscopes became commercially available in 1952 with a quite respectable 300-MHz effective bandwidth; significant sampling-gate device improvements have led to present-day bandwidth specifications as high as 18 GHz [34, 35, 36].

Today, a sampling oscilloscope may be purchased in analogous fashion to a real-time scope: a display main frame (either storage or standard phosphor CRT), a vertical *dual-channel* plug-in (with a variety of bandwidth-to-cost options), and a horizontal plug-in (with or without sweep delay, along with trigger-capability variations) comprising the system. Indeed, today, sampling plug-ins typically are interchangeable with real-time plug-ins in the same display main frame. As the cost and operational ease of sampling scopes have become steadily more competitive

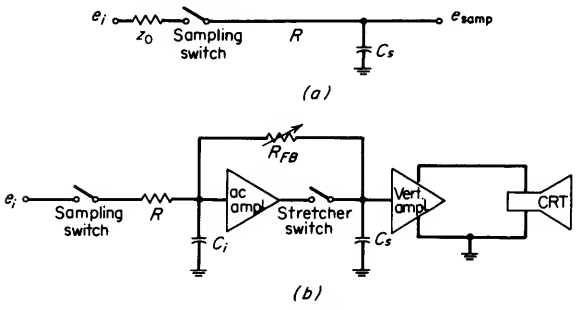


FIG 11-42 Sampling-vertical-amplifier block diagram.

with their real-time scope counterparts, it has become increasingly apparent that sampling oscillography is far more than just a replacement for a traveling-wave tube oscilloscope; the sampling scope has significant appeal as a general-purpose instrument.

Sampling Vertical. A basic sampling circuit is shown in Fig. 11-42*a*. It consists of a sampling switch, a series resistor, and a shunt capacitor to ground. At the instant that the switch is closed, the capacitor begins to charge. Since the switch is closed for a short time compared with the RC time constant, the capacitor will charge to only a small percentage of the actual signal amplitude ($e_{samp} = 0.05e_i$, for example).

Figure 11-42*b* shows a sampler circuit with a vertical amplifier and feedback circuit added. Sampling is accomplished by momentarily closing the sampling switch. The sampled voltage is transferred to the input capacitor C_i. This voltage is amplified and sent to the stretcher. The stretcher, or memory switch, is closed at the same time that the sampling gate is on, but remains closed for a much longer period of time. As a result, the stretcher capacitor has time to charge to the full voltage output of the ac amplifier. This voltage is applied to the vertical amplifier, where it is amplified sufficiently to drive the vertical deflection plates of the CRT. This new level is also fed back through a feedback attenuator, represented as R_{fb}, to the input capacitor. Gain of the ac amplifier and feedback is designed so that the voltage fed back to the input capacitor will be 100 percent of the sampled signal level; thus, when the next sample is taken, only changes from the previous level will be detected.

Example

Suppose that a +1-V signal is to be sampled. When the sample is taken, the input capacitor is charged to 5 percent of the input voltage, or 0.05 V. The amplifier increases this signal by a factor of 20 and applies 1 V to the stretcher

capacitor. The voltage on the stretcher capacitor is fed to the vertical amplifier and through the feedback loop to the input capacitor. By the time the oscilloscope is ready to take another sample, the input capacitor has charged up to 1 V in accordance with the feedback time constant. If the next sample is taken at the same input voltage level, no signal will be detected by the input capacitor and the dot displayed on the CRT will remain at the same vertical deflection.

Sampling Time Base. The time-base circuitry of a sampling oscilloscope differs greatly from that of a conventional oscilloscope. The function of the sampling time base is not only to move the dots across the screen in uniform increments of time, but also to generate a sampling command trigger for the vertical circuits.

Figure 11-43 shows an entire sampling system. The x-axis system consists of a sync circuit, time base, and horizontal amplifier. The sync circuit determines the sampling rate and establishes a reference point in time with respect to the signal. The time base generates both a timing ramp upon an output of the sync circuit and a staircase waveform which advances one step per sample. A coincidence of the timing ramp and the staircase level provides a sampling command to the sampler and stretcher switches. The horizontal amplifier builds the time-base signal to sufficient amplitude to drive the horizontal deflection plates.

A conventional time base produces a linear sawtooth sweep, to continuously move the beam horizontally across the CRT. The sampling time base also moves the beam across the screen, but usually not as a continuous movement. It positions the beam horizontally after a sample is taken and holds the beam at this location until the next sample is taken. The beam is then repositioned to a point slightly later in time on the CRT, where it again remains until the next sample. Thus, the time base is termed a *staircase-ramp generator*.

Sampler Operation. The basic elements of a sampling circuit are shown

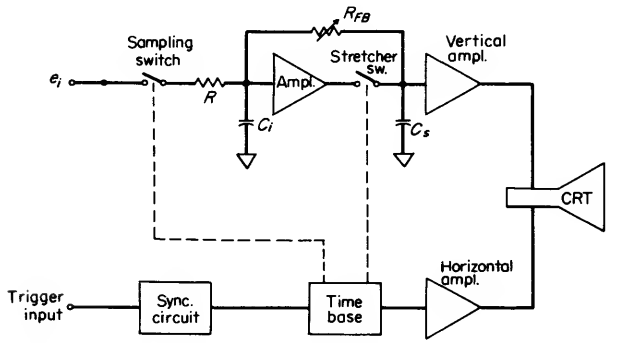


FIG 11-43 Sampling-scope block diagram.

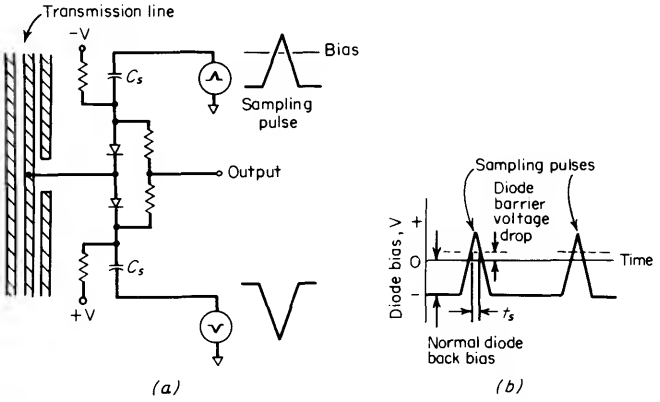


FIG 11-44 A two-diode feedthrough sampling circuit. Normally back-biased diodes are gated on by sampling pulses for short periods, allowing sampling capacitors C_s to acquire voltages proportional to the signal appearing on the transmission line.

in the idealized circuit of Fig. 11-42*a*. In this diagram, the system to be sampled is represented by a voltage generator e_i and an impedance Z_0, and the sampler consists of a sampling gate and a sampling capacitor C_s. When a sample is to be taken, the switch or gate is closed for a short period, which allows the sampling capacitor C_s to charge to some fraction of the input voltage e_i. The switch is then opened, with the sample of the input left stored on C_s.

If the voltage on C_s is assumed to be reset to zero before each new sample, a useful measure of the efficiency of the circuit is the sampling efficiency, defined as the ratio of the voltage on C_s after each sample e_{samp} to the input voltage e_i,

$$\eta = \frac{e_{\text{samp}}}{e_i} \tag{11-6-1}$$

Bandwidth of the sampler is defined as the frequency at which η is $1/\sqrt{2}$ times its dc or low-frequency value.

The time interval during which the act of sampling takes place must necessarily be finite. This sampling time and the charging-time constant of the sampling circuit limit the resolution and hence the rise time or bandwidth that can be achieved. These same factors also determine the transfer loss, or sampling efficiency, of the sampling circuit. Reduction of charging-time constant will improve both bandwidth and sampling efficiency, but for a given time constant, a decrease in sampling time will improve bandwidth only at the expense of sampling efficiency. These will now be examined in greater detail.

The manner in which sampling and charging times of the sampling circuit influence its response can be seen by considering the transient response to an impulse signal. The sampling circuit is considered to consist of a time-varying conductance $g(t)$ in series with a fixed capacitance C. The output of this circuit is the voltage on capacitor C just following sampling. This capacitor is assumed to discharge completely between samples.

Let it be assumed that the switching conductance function $g(t)$ is rectangular with amplitude G and duration T. The charging-time constant will then be defined as $\tau \equiv C/G$. The conventional transient response of this C and G circuit to a voltage impulse can be computed by usual means to yield

$$e_c = \frac{1}{\tau} e^{-t/\tau} \qquad \text{for } 0 < t \tag{11-6-2}$$

If an inpulse is now considered scanned by the sampling gate, with the stored value at termination of the gate taken as the sampled amplitude and time taken as the interval between the trailing edge of the gate and the impulse, then a similar wave will be reproduced except that it will be cut off at a time which corresponds to the gate length T.

$$e_{c1} = \frac{1}{\tau} e^{-t/\tau} \qquad \text{for } 0 < t < T \tag{11-6-3}$$

$$e_{c1} = 0 \qquad \text{for } t > T \tag{11-6-4}$$

Since an impulse has a flat frequency spectrum, the frequency transform of the impulse response will yield the frequency response of the system.

$$A(f) = \frac{1 - e^{(1+j2\pi f\tau)T/\tau}}{1 + j2\pi f\tau} \tag{11-6-5}$$

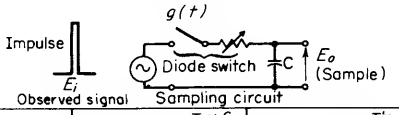
When $\tau \gg$ T or $\tau \ll$ T, the waveshapes degenerate to the forms given in columns 1 and 2 in Table 11-2. The responses for an assumed triangular $g(t)$ with $\tau \gg T$ are shown in column 3.

Sampling efficiency falls out of the above expression easily since it is equivalent to the dc gain of the system.

$$\eta = A(0) = 1 - e^{-T/\tau} \tag{11-6-6}$$

In practice the effective overall rise time of a sampling oscilloscope can be held close to the sampling duration limit. Subsequent amplifier rise time determines the speed with which the point is plotted and needs to be compatible only with maximum repetition rate capability [34]. Sampling gates have been developed greatly in recent years, along with models of their behavior. Reference 36 and its bibliography are especially good in this regard.

TABLE 11-2 Waveforms in a Sampling CRO



Diode switch conductance function $g(t)$	$T \ll C/G$	$T' \gg C/G$	$T \ll C/G$
Scope impulse response same as $g(t)$ if $T \ll C/G$, impulse response of charging CKT if $T' \gg C/G$			
Scope frequency response Spectrum of impulse response	$A(f) = \frac{\sin(\pi f T)}{\pi f T}$; Log freq; Log ampl.; Slope-1	$A(f) = \frac{G/C}{G/C + j2\pi f}$; Log freq.; Log ampl.; Slope-1	$A(f) = \frac{\sin^2 \frac{\pi f T}{2}}{(\frac{\pi f T}{2})^2}$; Log freq.; Log ampl.; Slope-2
Scope bandwidth 3 dB down	$\frac{0.442}{T}$	$\frac{G}{2\pi C}$	$\frac{0.636}{T}$
Scope step response integral of impulse response			
Scope rise time 10–90%	0.8 T	2.2 C/G	0.554 T

Sampling Heads. Several types of sampling gates have been employed. The first was a single-diode gate between Z_0 and C_s. The primary problems encountered were twofold: a "pedestal," or step, voltage occurred between the input and output because of the diode resistance, and a "kickout," or transfer of the sampling pulse into the system under test, was quite noticeable.

A four-diode bridge overcomes both problems reasonably well, provided the diode characteristics are well matched. A bridge combination keeps the total diode resistance in the path the same as for a single-diode sampler, since there are two series diodes in each parallel leg. The major difficulty with a four-diode gate is that the larger physical dimensions preclude maximum speed.

A two-diode sampler emphasizing speed was subsequently developed (Fig. 11-44*a*). This sampler is faster primarily because lead lengths between the signal sample point and the storage capacitor are shorter and hence subject to less degradation. The consequences of this design are twofold: Wideband noise is higher in this gate than in either of the first two described since, when on, the diode resistance is doubled; when they are off, "blow-by" or coupling around the sampling diodes due to stray capacitance is greater than with the four-diode gate.

The operation of the sampling gate is depicted in Fig. 11-44*b*. The diodes are off until a sampling pulse, which is a brief-duration balanced

pulse applied differentially across the two-diode (or four-diode) gate, is received. During time t_s, the diode is forward biased and any charge level on the signal line is conveyed to C_s and stored there. Very fast sampling diodes are used in order to respond to turn-on and turn-off signals at the proper time. The total sample time t_s is frequently in the tens of picoseconds in current instrumentation.

A recent development (Fig. 11-45) uses a somewhat different approach to a sampling gate. It may be represented as a pair of diodes, separated by a known delay line section, replacing each diode in the two-diode gate. This arrangement allows a wide strobe or pulse gate, with a narrow time window for sampling defined by the different responses of the two diodes due to the line. Consequently the sampled signal is defined as the charge trapped on the delay line section during the short time window when one diode is turned on and the other is off. Potentially this may be the fastest gate technique yet developed [37, 38].

Triggered and Random Sampling. The major advantage of the sampling technique is that only the sampling head must work at the very high speeds of the signal to be measured, while the rest of the circuits may work at relatively low speeds. However, the delays of the time-base circuits have to be compensated for in order to view the leading edge of a signal. This requires, then, a delay line in the vertical signal path which must be in front of the sampling gate (in contrast with the conventional oscilloscope where it may be anywhere in the vertical amplifier).

If the delay line is inserted in the signal path, it must transmit the full signal bandwidth. This method is often found in sampling units

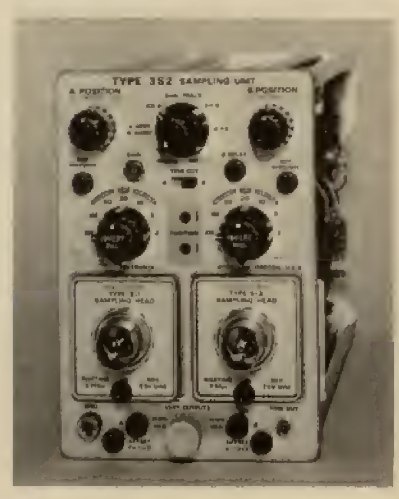


FIG 11-45 A vertical-amplifier sampling plug-in with interchangeable heads. (*Tektronix, Inc.*)

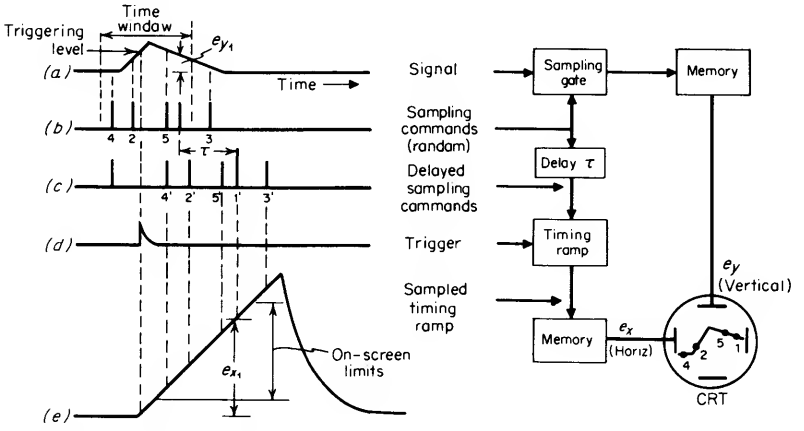


FIG 11-46 Random sampling: derivation of x and y deflection signals.

having an *internal* trigger mode. Another method using a delay line is to insert it between a clock and the stimulus pulse generator that produces the signal to be measured. Then the trigger input is derived from the clock, and after suitable delay the pulse generator is triggered by the same clock.

A wholly different approach, *random sampling*, avoids a delay line altogether. The advantages are several with this technique: The low impedance (50 Ω) of a delay-line input on a standard sampling unit is avoided, the inherent distortions and bandwidth limitation of the line are eliminated, and display jitter can be reduced further than is typically found in delay-line-triggered systems. Offsetting these advantages is the fact that random sampling requires a relatively high trigger repetition rate to obtain adequate samples in a given time interval for display. A camera or a storage CRT are of value in reducing this disadvantage.

Random-sampling operation is a two-step process. The first step is to develop a useful *sampling distribution*, a high density of samples during the time window when a signal is expected and a low density at other times. After this is done, two analog signals e_x and e_y are derived to represent the x and y coordinates of any particular sample.

The y (or vertical) coordinate is obtained by the same sample-and-hold process used in a conventional sampling oscilloscope. The x (or horizontal) coordinate is derived as shown in Fig. 11-46. The figure illustrates five randomly placed samples of the signal. Each of these samples was taken on a successive repetition of the signal.

The y component e_y of the first sample is held and subsequently used to

position the CRT spot vertically. The sampling command which asked for the first sample is then delayed by a fixed time interval τ. This delayed sampling command 1′ is used to sample a timing ramp which was started by *trigger recognition* along the input signal at t_0. The resulting sample e_x is held and later used to position the CRT spot horizontally.

By repeating this process, subsequent samples supply both vertical and horizontal information to deflect the CRT beam from dot to dot and thus construct a display of the signal from those samples which fall within the time window.

Note that increasing τ will provide more lead time in the display for a given signal transition. Such an increase in τ requires a time shift of the sampling distribution to an earlier point in time in order to collect usable samples for the display [39, 40, 41].

11-7 Special-purpose Oscilloscopes

A great variety of measurements are made with special-purpose instruments using basic oscilloscope techniques. Within the class of time-base display instruments, there are *digital-readout oscilloscopes*, *time-domain reflectometers*, *television video monitors*, and *medical heart-rate monitors*. If frequency instead of time is displayed on the x axis, the unit may be called a *swept-frequency indicator* or a *spectrum analyzer*. Often, the x and y axes are simply used for a linear plot of any variable, much like an xy recorder display. *Phase plotting* is a very common requirement, as are graphic plotting and alphanumeric readout in *computer terminal displays*. Such displays are also used in component test instruments such as *transistor curve tracers*.

Digital Readout. Capturing and displaying an accurate rendition of a signal are two of the three important functions of an oscilloscope; the third function, analysis, is usually done by the operator by correlating the displayed signal with the control panel settings and the CRT graticule. The digital-readout oscilloscope seeks to improve this function both by increasing the convenience of obtaining an answer and by increasing the accuracy of the answer [42].

Units available at the present time are capable of measuring between any two time references, or any two voltage references. Thus a rise time, fall time, or pulse width may be determined while using standard definitions (10 to 90 percent on t_r, t_f; 50 percent on pulse width) or any two variable levels. A reference zone establishing 0 and 100 percent levels must first be located, either by predetermined levels or adjustment by the operator. These levels then serve as reference for any measurement of vertical amplitude, as well as limits of comparison for threshold detectors determining the start and stop points for any time measurement.

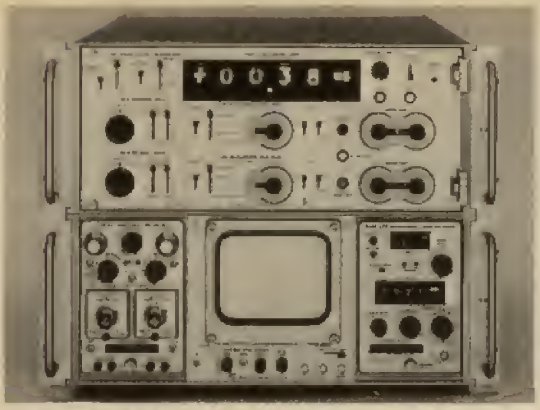


FIG 11-47 A digital-readout oscilloscope system. (*Tektronix, Inc.*)

The biggest use of digital oscilloscopes today is in automated production-line testing, such as in the semiconductor manufacturing industry, where numerous go–no-go tests of a simple waveform-analysis nature must be repeated on each component. The digital-readout scope, especially when provided with programmable controls, is uniquely suited to this application (Fig. 11-47). As analog-to-digital techniques at oscilloscope test frequencies become cheaper, and large-scale integrated circuits become more common, it is not unlikely that the oscilloscope with digital readout will become more accepted as a general-purpose instrument, just as the digital-readout voltmeter has become virtually as common for laboratory analysis as an analog voltmeter with a D'Arsonval meter readout.

Time-domain Reflectometry. Inserting a pulse of energy and monitoring the time-domain reflection of that energy by an oscilloscope system comprises the time-domain reflectometry technique. Essentially a closed-loop radar system in principle, the method has become extremely valuable in transmission-line analysis and design in recent years, with the advent of special instrumentation designed for the purpose.

A typical time-domain reflectometry unit inserts a voltage step to be propagated down a transmission line under evaluation and displays on a CRT both the incident and reflected voltage waves as measured at the insertion point. Analyzing the time delay, magnitude, and shape of the reflected waveform permits determination of location and nature of the impedance variations in the transmission line.

The relative amplitude of the reflected signal correlates with the impedance of a discontinuity. The distance of the discontinuity from the signal input can be determined by the time-domain pulse separation

between incident and reflected wave. The distance to a discontinuity is

$$d = \frac{ct_0}{2(\epsilon)^{1/2}} \tag{11-7-1}$$

where c is the speed of light, ϵ is the relative dielectric constant of the transmission line, and t_0 is the elapsed time between generated and reflected pulses. If the distance between two similar discontinuities is

$$d_{1-2} = \frac{c(t_1 - t_2)}{2(\epsilon)^{1/2}}$$

it becomes impossible to resolve their separation when $t_1 - t_2$ becomes less than one-half the time-domain-reflectometer pulse-system rise time. With a time-domain reflectometry system composed of a sampling oscilloscope and a fast-rise pulse generator, distance resolution between discontinuities as close as 0.25 in. is achievable.

A feed-through sampling head may be employed in either of two ways to determine transmission-line characteristics (Fig. 11-48). In the reflectometer mode, a pulse input is fed through the sampler to a system under test, and the display is of both the incident and reflected waveforms as seen by the feed-through sampler. The transmission mode displays only the transmitted wave from the system under test and is unable to distinguish time separations between discontinuities, as can reflectometry, but it can display the total time-domain transmission characteristic of a transmission line. The transmission mode reveals the equivalent time-domain picture of a system transmission characteristic that swept-frequency generator and monitor techniques provide for the frequency domain.

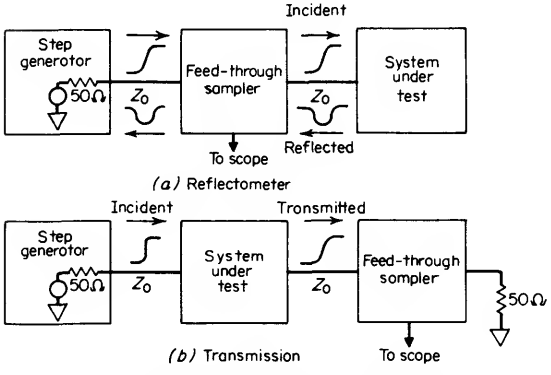


FIG 11-48 Time-domain reflectometry techniques.

The classic measurement of transmission-line characteristics by reflection, of course, involves inserting a continuous sine wave into a system and measuring the maximum and minimum amplitudes of the standing waves resulting from line discontinuities. From such measurements at a given frequency, a standing-wave ratio is calculated. Single-frequency standing-wave-ratio techniques fail to reveal whether a given discontinuity is generating a reflection of proper magnitude and phase to cancel one from a different discontinuity; thus standing-wave-ratio measurements must be made at many frequencies, a rather time-consuming and tedious task. Swept-frequency reflectometers are available today to facilitate the determination of standing-wave ratio.

Such units give a magnitude reflection coefficient as a function of frequency, but no phase or location information is available.

Time-domain reflectometry is a powerful design and analysis tool unmatched by any standing-wave-ratio technique, for it truly separates the location and magnitude of a number of minor discontinuities contributing to the aggregate effect. Perhaps more important, the type of mismatch at each discontinuity (R, L, or C or a combination) is readily determined. Figure 11-49 illustrates three cases of simple discontinuities as they would be displayed with a reflectometry mode measurement on a system calibrated to 50 Ω. The reflection coefficient may be measured (with respect to the unit step) on the scope graticule and the purely resistive load R_L calculated directly. In the reactive cases, τ may be measured and L or C calculated. More complex terminations, both shunt and series, may also be evaluated easily, as shown in Fig. 11-50 [43].

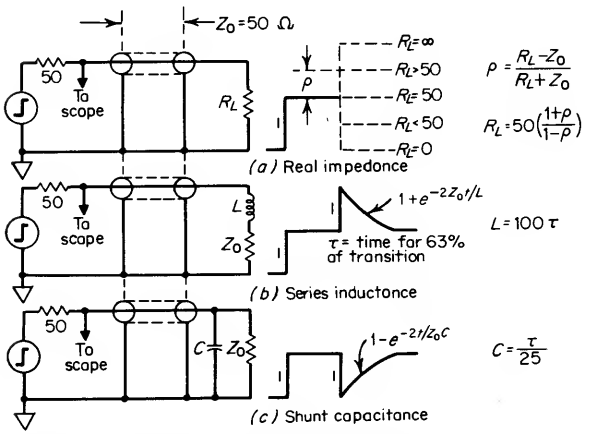


FIG 11-49 Resistive and reactive impedance determination with time-domain reflectometry.

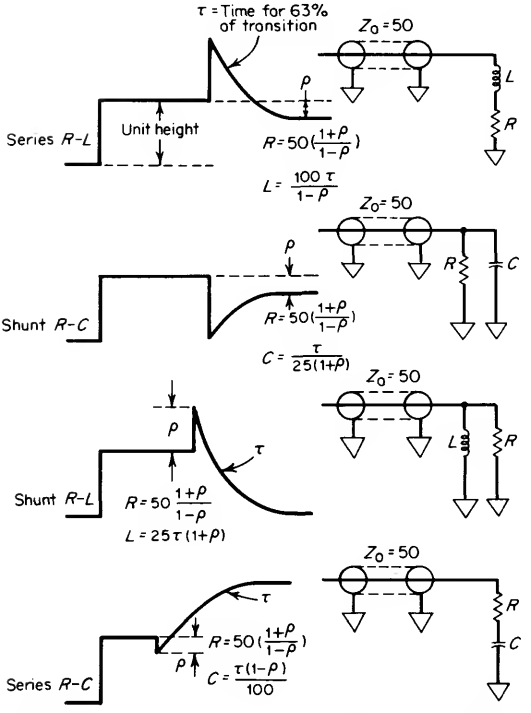


FIG 11-50 Analysis of complex line and load characteristics with time-domain reflectometry.

Of prime value in electrical cable and transmission-line testing, time-domain reflectometry has also been applied to many other situations where location of moving materials in conduits is unknown. For example, in oil-refinery cracking towers, time-domain reflectometry can quickly indicate the level of each layer of oil product (kerosene, gasoline, butane, etc.). Studies of the effect of light-source stimuli on cockroach behavior were rendered possible by using time-domain reflectometry to watch the movements of the insects in a closed system. It is used in England to monitor soup cauldrons for proper soup homogeneity. Checking thermocouple characteristics after implantation in atomic piles is another difficult job rendered practical by time-domain reflectometry [44].

The *xy* Plotters. An external horizontal input which omits the time-base generator is a common provision on oscilloscopes (Fig. 11-1). This provision allows the user to plot another variable instead of time on the x axis. For example, spectrum analyzers plot frequency, curve tracers plot voltage, and vector cardiology instruments plot phase vectors.

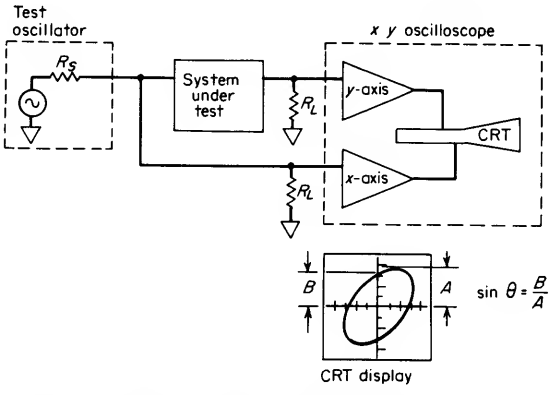


FIG 11-51 Phase-measurement block diagram. The x- and y-axis amplifiers should have matched phase characteristics for the frequencies of interest. Note that R_L for each axis should be the same to avoid introducing a phase-shift difference from the scope input RC.

If the x and y amplifier characteristics of the oscilloscope are nearly identical, a quite useful phase-shift analysis is possible with it. Figure 11-51 gives a block diagram of the measurement technique, from which it may be seen that identical x and y amplifiers in the oscilloscope will allow determination of the difference in both amplitude and phase between the input and output of the circuit under test. Cathode-ray-tube displays of phase measurements for three different circuits are shown in Fig. 11-52.

Typically, oscilloscopes with a horizontal input to permit xy phase

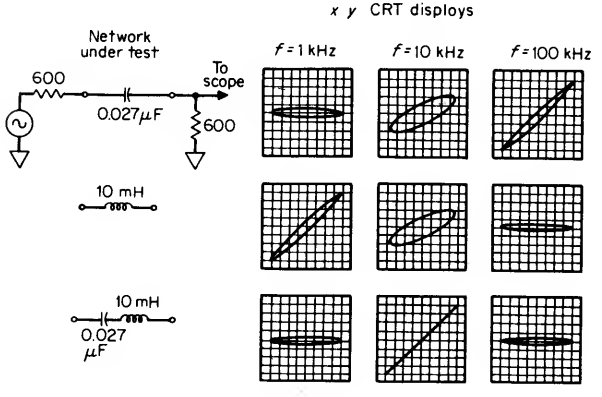


FIG 11-52 Phase-measurement displays.

plotting are capable of accuracies to within 1° of phase shift from dc to 50 kHz. This is a useful frequency range for phase measurements of such phenomena as *BH* curves of transformer materials and medical vector cardiology. For higher-frequency phase-shift measurements, the delay line in the vertical amplifier must be eliminated and such effects as beam transit time between the vertical and horizontal plates of the CRT must be compensated for. Units with less than 1° of phase shift between x and y inputs to 10 MHz have been commercially available on special request; this corresponds to a time delay difference of less than 0.3 nsec between the vertical and horizontal system.

Frequency Domain. Although time-domain analysis has become very popular because of the measurement capability afforded by the oscilloscope, there remain many classes of measurements for which frequency spectrum information is necessary. In recent years, the oscilloscope display and *frequency sweeper* have been combined into an extremely versatile instrument called a *spectrum analyzer*. See Chap. 16 and Refs. 45 and 46.

Medical Displays. Variations of the standard oscilloscope are used in many medical applications. Patient monitoring in surgery, coronary care, intensive care, and recovery areas is greatly facilitated by several scope types of display. Various physiologic functions such as electrocardiographic waveforms, arterial and venous pressures, and respiration rates are monitored continuously by time-base scope monitors, which often include alarm systems to indicate abnormal phenomena. Both clinics and research laboratories rely upon such diagnostic instruments as vectorcardiographs, electromyographs, and diagnostic sounders, all of which use oscilloscope display techniques.

The surgery-room monitor is usually a large-screen multichannel oscilloscope with a slow-speed time base appropriate for physiologic monitoring. The patient-care units found alongside beds and in nurses' stations are similar instruments in many respects, although they are often only single-channel displays coupled with meter readouts. Either type may incorporate threshold detection for an automatic alarm system whenever a monitored function such as a heartbeat rate exceeds desired limits (Fig. 11-53). Monitors of each of the standard physiologic functions—electrocardiographic; heart rate; systolic, diastolic, or mean blood pressures—are available in various instruments [47].

The vector cardiograph displays a vector, or phase plot, of the heart beat phenomena instead of the y-axis-versus-time plot of the standard scope monitor. Some units are able to display both parameters, either simultaneously or with operator selection of the mode of display [48]. Standard dual-beam oscilloscopes may also be used for this purpose.

Two units available either as standard equipment instrumentation or plug-ins for a standard oscilloscope main frame are the electromyograph

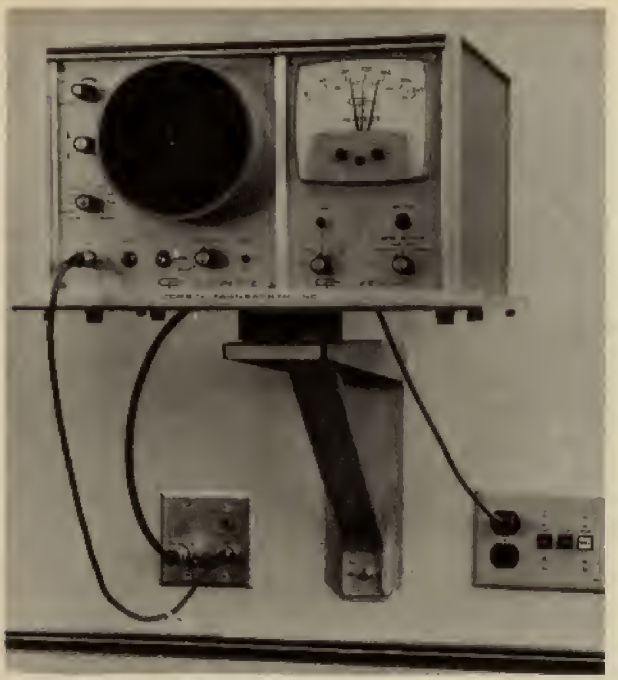


FIG 11-53 A heart-rate medical monitor. (*Smith Kline Instruments, Inc.*)

and the diagnostic sounder. The former is used primarily to monitor nerve conduction and the electrical activity of muscle tissue. A variable-persistence CRT is especially useful with this instrument, for it allows easy display of nerve-conduction velocity measurements which are carried out at low stimulus repetition rates.

The diagnostic sounder relies upon ultrasonic techniques very similar to the time-domain reflectometry technique previously discussed. Four stimulus frequencies are typically provided between 1 and 10 MHz, to facilitate study of wave transmission and reflection in various portions of the body. The system transmits short pulses at the stimulus frequency into a transducer in firm contact with the body (aided frequently by a coupling medium such as transmission gel). When the pulses encounter a tissue interface or internal structure with a different density or elasticity from the surrounding tissue, an echo pulse is reflected to the transducer and displayed on the CRT. With a calibrated display, the depth of the anomalous tissue may be read directly. Some of the more useful areas of study with this instrument include determination of the brain midline,

observing heart-valve motion, detecting pericardial effusion, and locating either foreign material or tumors within the body [49].

Computer Information Display. Computer terminals today feature the oscilloscope in a relatively new and intriguing field. Since the advent of computers, emphasis has been placed upon the communication link between men and machine. Machine language (computer) was the first point of interconnection; assembly language (Fortran and Algol) followed shortly to help the man communicate more easily. In recent years, both voice and hand communication in conversational language (English) has been the goal. The CRT display has lent itself well to this increased ease of computer communication, and many installations today are using CRT presentations for information display. Large displays from CRTs with from 12- to 27-in. diagonal are more suitable in this field than the traditional 5-in. oscilloscope display [50, 51, 52].

Computer terminals began with modified television displays having raster-scan display circuitry. A need for densely packed character formats with bright-contrast ratios has led to random-access printout formats. As higher computer clock rates are used, magnetic displays often become display limited because of the relatively slow xy deflection rate of the magnetically deflected CRT and the amplifiers.

Magnetic displays have benefited greatly in recent years from refinements in transistors and deflection yokes. High-power, high-current devices are paralleled (occasionally more than 60 devices in parallel are in the output deflection stage) to decrease the jump-scan time (the deflection time between diagonal corners of the CRT), while ultralinear yokes and high-frequency sensor and feedback circuits are combined to make the response as linear as possible. Such highly developed amplifiers make the random-access magnetic display unit expensive.

Another approach includes both magnetic and electrostatic deflection in a combination CRT, employing the magnetic deflection for beam movements larger than 1 in. and electrostatic deflection for small movements. This enables one to write many characters in a 1-in. square very rapidly and then go relatively slowly to the next square to repeat the fast sequence. Many more characters may be written this way than with a conventional magnetic display unit, yet only a small electrostatic deflection amplifier and a simple magnetic drive amplifier are required. The CRT cost offsets part of the advantage, and the resulting computer constraints in display programming are also disadvantageous.

Electrostatic displays historically have not found much use in computer terminals. Compared with magnetic deflection tubes, electrostatic deflection CRTs typically have very low beam-scan angles, which means that the CRT length must be much longer to obtain the same display area. Also, since voltage rather than current is the deflection input in

electrostatic tubes, transistor technology has been more adaptable for amplifiers to deflect magnetic CRTs. A typical electrostatic tube may be 30 in. long, with deflection factors of 75 to 100 V/in., to achieve an 8 × 10 in. display area. Resulting deflection amplifiers have mostly been of transmitter-tube designs, with attendant problems of high power.

Expansion-mesh technology applied to large-screen electrostatic tubes to obtain wide-angle deflection has permitted both increased deflection sensitivity and shorter tube length. The major disadvantage of such a design is beam expansion, which results in a larger spot size. Nonetheless, such CRTs are finding favor in some computer applications where high speed of deflection and reasonable display cost are necessary. Multichannel multiplexing is especially facilitated by such a display [53].

Storage displays also meet the requirement for information of high density in the computer terminal. This approach features a very low refresh rate (perhaps one frame in 2 min) combined with a standard television writing rate for magnetic deflection, rather than a standard refresh rate (30 to 60 frames per second) with high-speed writing. One unit using phosphor-storage techniques is presently available in a 12-in. version (Fig. 11-54), and it is beginning to find widespread application for certain types of computer terminals. Such a display lends itself well to data transmission by telephone between the terminal and the control computer, since the storage time is compatible with low data rate of telephone-line grade of transmission. One disadvantage of the method is the lack of efficient interaction for the operator by means of light pen or other link, since selective erasure and modification are difficult. More serious for picture transmission is the lack of halftones, since the only storage unit available with large-screen capability today uses phosphor storage [54, 55].

Curve Tracers. Curve tracers which use oscilloscope techniques for testing active devices such as vacuum tubes, transistors, and integrated circuits are common test instruments. The display is usually a voltage-versus-current graph of the dynamic characteristics of the device under test with voltage displayed on the x axis and current on the y axis. A step generator may be used to advance a third variable so that a family of curves is displayed. Controls are available to vary the dynamic test conditions over a considerable range, enough that low-level V_{be} characteristics and collector saturation resistance ($r_{c\,sat}$) can be examined as well as BV_{ceo} for a device requiring several hundred volts. Common parameters monitored include breakdown voltage, dynamic resistance, alpha and beta current gain, gain linearity, and leakage currents [56].

As with most oscilloscopes, curve tracers have a large number of controls to allow a great flexibility in measurement capability. This is an advantage for the skilled operator, but a handicap for the occasional user. A

FIG 11-54 A computer terminal using a phosphor-storage CRT display. (*Tektronix, Inc.*)

recent improvement in curve-tracer convenience provides range setting information on the display panel so that the knob settings may be more readily noted (Fig. 11-55). While its major advantage is in locating the scale information in a more convenient place for the user, additional benefits include the ease of photographing both the analog picture and its range coordinates, plus provision for some simple derived parameters, such as the beta division number provided on the illustrated instrument [57].

Miscellaneous. There are numerous other variations of oscilloscope types of instrumentation. For example, simulated three-dimensional projections are available with a *scenoscope*, nuclear-event monitoring is provided by a *pulse-height analyzer*, aircraft or other noise monitoring is observed on a *loudness analyzer*, and automotive tune-ups are improved by an *engine analyzer*. *Television video monitors* are scopes that use a time-base display with a TV line sync and a specially shaped vertical passband; *television picture monitors* are essentially high-resolution television sets complete with raster-scan deflection circuits.

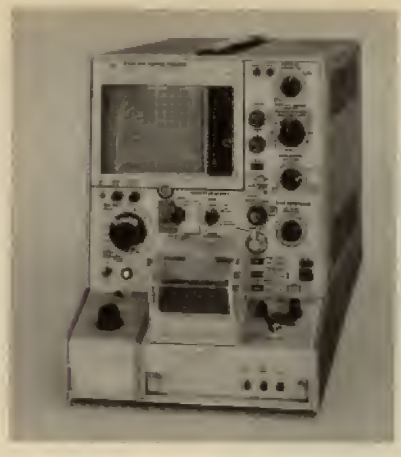


FIG 11-55 A curve tracer with alphanumeric display. (*Tektronix, Inc.*)

11-8 Accessories

A large number of accessory items is available to increase the usefulness of the oscilloscope for special applications. Transducers for signal capture and recorders for a retained hard copy of the displayed phenomena are the two most important accessories. Other items include CRT filters or viewing hoods to reduce light reflections or to enhance contrast, scope carts for ease in moving larger oscilloscopes, carrying cases for portable oscilloscopes, and special servicing equipment such as storage cabinets (with power) and plug-in extenders. The following discussion will be restricted to transducers and recorders.

The transducer is the coupling device between the quantity under test and the oscilloscope vertical amplifier. It must satisfy several stringent requirements: The signal to be observed must not be distorted appreciably by the transducer, the output of the transducer must be compatible with the oscilloscope input, and the transducer must be small enough and sturdy enough to allow reliable use by the operator in awkward places. These criteria are met to varying degrees by the different transducers available.

The simplest transducer is a voltage probe, with input and output impedance and gain (or attenuation) specifications. A wire from the scope input to the circuit under test will serve as a probe with 1:1 gain ratio, and input and output resistance is that of the scope input (typically 1 MΩ). This technique is often adequate for low-frequency measurement, but it suffers from both noise pickup of any stray signals (such as 60-Hz power lines) and increased capacitance and inductance loading on the circuit under test.

A coaxial cable may be used to shield the probe wire from stray signals, but it adds significantly to input shunt capacitance. To overcome this,

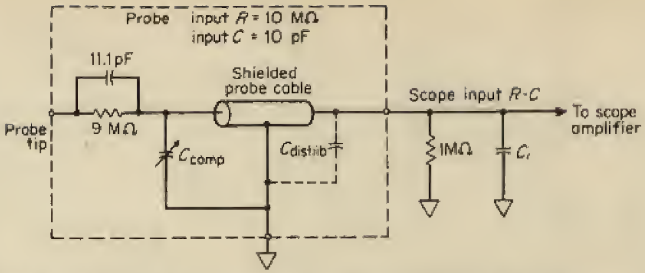


FIG 11-56 Compensated 10:1 voltage-divider probe schematic. The compensating capacitor C_{comp} is adjusted for flat-top response on a square-wave signal. The condition of compensation is that $C_{comp} + C_{distrib} + C_i = 100$ pF.

the compensated divider probe has been developed (Fig. 11-56). Such a probe, available usually with a 10:1 or 50:1 voltage attenuation ratio, presents a higher input resistance and lower shunt capacitance to the circuit under test than the oscilloscope input itself. The probe is capable of attenuating all frequencies equally when the adjustable compensating capacitor is properly adjusted, just as is the compensated attenuator discussed previously. This type of probe is by far the most common coupling device used for oscilloscope measurements (Fig. 11-57). Probe

FIG 11-57 A compensated 10:1 divider probe at lower right, 1:1 probe at upper right. Other accessories include camera bezel adapters and cables. (*Hewlett-Packard Company.*)

compensation is readily accomplished by displaying the calibrator waveform and adjusting for a flat pulse response (Fig. 11-58). Most compensated-divider probes are capable of probing signals to 600-V dc without damage, and special high-voltage units with attenuation ratios to 1,000:1 are capable of probing to 40-kV levels.

Active voltage probes use an internal amplifier to achieve a 1:1 gain ratio while providing the attributes of higher input impedance and shielding of the compensated attenuator probe. Their chief limitation is a relatively low dc dynamic range (± 0.5 to ± 5 V) since they use an active device such as an FET at the probe input.

Bandwidth specifications for probes are somewhat different from those for oscilloscopes. The typical method is to specify the probe rise-time capability and then calculate the total system capability from probe tip to CRT display with the following equations:

$$t^2_{r\text{system}} = t^2_{r\text{probe}} + t^2_{r\text{scope}} \tag{11-8-1}$$

$$BW_{\text{system}} = \frac{0.35}{t_{r\text{system}}} \tag{11-8-2}$$

At higher frequencies, the probe capacitance becomes the most significant impedance loading factor. Consequently, a high-speed system should include a probe with a very low input capacitance. Indeed, in both real-time and sampling systems over 100 MHz, there are 50-Ω probe

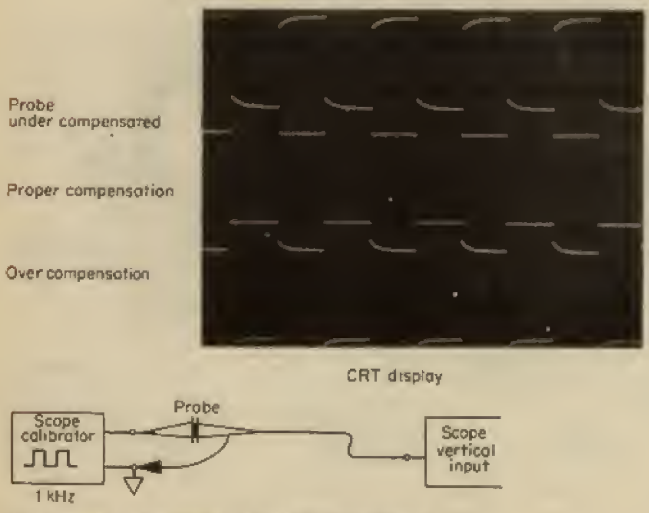


FIG 11-58 Calibration of probe compensation.

systems as well as high-resistance probes with low input capacitance. The lower-frequency restrictions of a 50-Ω system are severe, but at higher frequencies, the accuracy may be much better for certain measurements [58, 59].

Sampling probes may use either a feed-through or a terminated 50-Ω line technique. Usually the sampling probe is an integral portion of a high-impedance sampling-scope vertical amplifier; so the probe is not necessarily an accessory. Terminations and resistive divider networks are available for sampling probes, as are blocking capacitors. Special high-speed pulse generators (for sampling or time-domain reflectometry work), filter networks, and trigger countdown units are among the other sampling-scope accessory items [58, 59].

Current probes to convert a current signal into a voltage signal are also available for use with oscilloscopes. Some units rely upon transformer coupling for the transducer action and are thus restricted to ac measurements. Others use magnetic field sensing (such as a Hall effect transducer) to extend the current-probe capability to dc [60, 61]. Current-probe amplifiers are available to increase the transducer gain by 50 or more from the typical 1 mA to 1 mV sensitivity.

Other transducers used with oscilloscopes include piezoelectric crystals, strain gages, and medical transducers such as ear or finger plethysmographs and pulse-wave electrodes.

Hard Copy. There are two common ways to provide a hard-copy print of an oscilloscope display for a historical record. Film photography is the more flexible and popular method for the occasional user, while *xy* recorders have several advantages for certain measurements which justify their use.

Oscilloscope cameras are accurate and convenient for the recording of the scope display under virtually any display condition, whether it is a repetitive sampling picture, a multiplexed series of signals, or a very high speed single-event transient. The camera is mounted directly over the CRT, and the full CRT display and graticule are recorded (Fig. 11-59). The use of transfer film allows development of the picture in a few seconds, and a hot stylus (such as a hot soldering iron tip) may be used to etch the *xy* coordinates directly onto the film. Standard 4 × 5 in. or 2¼ × 3¼ in. film negatives may be made with different camera backs if a number of prints are desired.

Shutter speed, focal length, object-to-image ratio, and lens aperture are typically all adjustable to accommodate a wide variety of display conditions. A viewing window is usually provided for the operator to see the CRT and adjust the display for proper photography. This also facilitates multiple-image photography. If the scope CRT has an internal graticule, one of two means of photographing the graticule is provided.

FIG 11-59 Camera mounted on oscilloscope. (*Hewlett-Packard Company.*)

The CRT may have a back-lighted or edge-lighted faceplate which photographs as a white graticule and white phosphor trace against a dark background (Fig. 11-60*a*.) Inclusion of an ultraviolet light in the camera or a graticule flood gun in the CRT permits low-level excitation of the background phosphor, which results in the graticule's photographing as a black trace, the background in gray, and the trace in white (Fig. 11-60*b*).

Choice of a camera and film system varies with the requirement almost as much as the scope itself. If ultrahigh writing speed is desired in order to record single-shot phenomena, the selection would probably include an f/1.3 lens, Polaroid film type 410 with an American National Standards film equivalent exposure index of 10,000, a 1.0:0.5 object-to-image ratio, and a prefogging capability to enhance the film sensitivity. For standard photography, an f/1.9 lens, film with a 3,000 American National Standards equivalent, and a 1:1 object-to-image ratio are more satisfactory for both cost and usability. Film types and characteristics, camera capability, and mating with the proper oscilloscope for a given measurement are subjects for which most manufacturers will provide individual suggestions [62, 63].

A paper hard copy from an *xy* recorder has two advantages over a film photograph hard copy: It uses a cheaper material than film (of interest if

great quantities of records are required), and it is an easier copy to work with (whether the user is writing additional information on it, pasting it into a notebook, photocopying it for distribution, or reading the signal resolution). The disadvantages of *xy* recording are several: The wide disparity between the electron-beam deflection rate and the recorder-stylus deflection rate requires a frequency transducer between the two devices (which effectively requires a sampling transducer and thus a repetitive signal), the difficulty of obtaining a copy may well offset the material cost savings, and multichannel displays are especially difficult to plot without specialized transducers and recorders. Recorder advances are reducing these problems, but the camera does enjoy a flexibility and convenience for most users which the recorder cannot match.

Since sampling scopes already employ a frequency transducer, it is not uncommon to find 1-V full-scale x and y outputs provided with sampling systems. These outputs are compatible with most *xy* recorders available commercially.

A scope plug-in recorder has been available to convert any repetitive

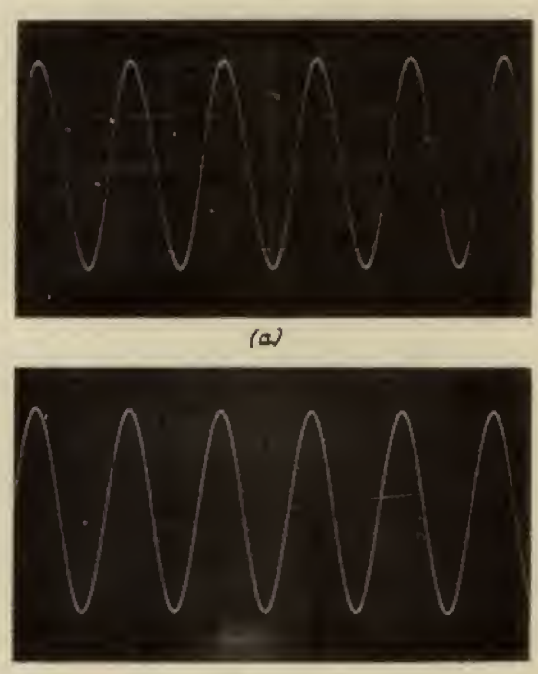


FIG 11-60 Graticule illumination.

signal display to a small paper copy by use of a display scanner, sampling, and a galvanometer-actuated stylus. It can copy in a few seconds any repetitive signal to 30-MHz bandwidth displayed on the scope. Because of the nonrepetitive nature of noise, such recorder techniques are especially suited for signal averaging to obtain clear signals in the presence of considerable noise [64, 65].

CITED REFERENCES

1. Guillemin, E.: "Introductory Circuit Theory," John Wiley & Sons, Inc., New York, 1953.
2. Scott, R. E.: "Linear Circuits—Time-domain Analysis, Part 1," Addison-Wesley Publishing Company, Inc., Reading, Mass., 1960.
3. Churchill, R. V.: "Modern Operational Mathematics in Engineering," 2d ed., McGraw-Hill Book Company, New York, 1958.
4. Bakish, R.: "Electron Beam Technology," John Wiley & Sons, Inc., New York, 1962.
5. Parr, G., and O. H. Davie: "The Cathode-ray Tube and Its Applications," 3d ed., Reinhold Publishing Corporation, New York, 1959.
6. Spangenberg, K. R.: "Vacuum Tubes," McGraw-Hill Book Company, New York, 1948.
7. "Cathode Ray Tubes," Tektronix, Inc., Beaverton, Ore., 1967.
8. Moss, H.: The Electron Gun of the Cathode Ray Tube, *J. Brit. Inst. Radio Engrs.*, Pt. 1, January, 1945; Pt. 2, December, 1945.
9. Luxenberg, H. R., and R. L. Kuehn: "Display Systems Engineering," Inter-Univ. Electron. Series, vol. 5, McGraw-Hill Book Company, New York, 1968.
10. Soller, T., M. A. Starr, and G. E. Valley, Jr. (eds.): "Cathode-ray Tube Displays," MIT Rad. Lab. Series, vol. 22, McGraw-Hill Book Company, New York, 1948.
11. Anderson, R. H.: A Simplified Direct-viewing Bistable Storage Tube, *IEEE Trans. Electron Devices*, vol. ET-14, no. 12, December, 1967.
12. Chance, B., V. Hughes, E. MacNichol, D. Sayre, and F. Williams: "Waveforms," MIT Rad. Lab. Series, vol. 19, McGraw-Hill Book Company, New York, 1948.
13. Knoll, M., and B. Kazan: "Storage Tubes and Their Basic Principles," John Wiley & Sons, Inc., New York, 1952.
14. "Storage Cathode-ray Tubes and Circuits," 2d ed., Tektronix, Inc., Beaverton, Ore., 1968.
15. Waters, W. M.: Electronic Half-tone Image Recording Technique, *IEEE Proc.*, vol. 54, pp. 319–320, February, 1966.
16. Kolar, R. H.: Variable Persistance Increases Oscilloscope's Versatility, *Electronics*, vol. 38, Nov. 29, 1965.
17. Harsh, M. D.: Display and Storage Tubes, *Electronic Inds. Tele-Tech*, vol. 24, April, 1966.
18. Yaggy, L. S., and N. J. Koda: A Versatile, High-performance Scan Converter Storage Tube, *Proc. 8th Natl. Symp., Soc. Inform. Display*, May, 1967.
19. Czech, J.: "Oscilloscope Measuring Technique; Principles and Applications of Modern Cathode Ray Oscilloscopes," 2d ed., transl. from German, Philips Technical Library, The Netherlands, 1965.
20. Rider, J. F., and S. D. Uslan: "Encyclopedia on Cathode Ray Oscilloscopes and Their Uses," 2d ed., John F. Rider, Publisher, Inc., New York, 1959.

21. Valley, G. E., Jr., and H. Wallman (eds.): "Vacuum Tube Amplifiers," MIT Rad. Lab. Series, vol. 18, McGraw-Hill Book Company, New York, 1948.
22. Pettit, J. M., and M. M. McWhorter: "Electronic Amplifier Circuits," McGraw-Hill Book Company, New York, 1961.
23. Middlebrook, R. D.: "Differential Amplifiers: Their Analysis and Their Applications in Transistor DC Amplifiers," John Wiley & Sons, Inc., New York, 1963.
24. Hegeman, B.: How to Reduce Common Mode Signals in Oscilloscope Measurement, *IEEE, Circuit Design Eng.*, February, 1966.
25. Horn, J. J.: Differential Comparator Extends Measurement Accuracy, *Electronic Design*, vol. 13, Oct. 25, 1965.
26. Rheinfelder, W.: "Design of Low-noise Transistor Input Circuits," Hayden Book Company, Inc., New York, 1964.
27. Pettit, J. M.: "Electronic Switching, Timing, and Pulse Circuits," McGraw-Hill Book Company, New York, 1959.
28. Millman, J., and H. Taub: "Pulse, Digital, and Switching Waveforms," McGraw-Hill Book Company, New York, 1965.
29. Understanding Delaying Sweep, *Service Scope*, no. 50, *Tektronix, Inc.*, June, 1968.
30. Zimmerman, H. A.: Pseudo-Schmitt Eliminates Uncertainties in Trigger Logic, *Electronic Design*, vol. 13, Nov. 22, 1965.
31. Grein, W.: Real-time Oscilloscope Triggering, *Elec. Design. News*, Nov. 22, 1967.
32. Lewis, F. D., and R. M. Frazier: "Distributed-parameter Variable Delay Lines Using Skewed Turns for Delay Equalization," *Proc. IRE*, vol. 45, pp. 196–204, February, 1957.
33. Siegel, F. G.: "A New DC-50+ MHz Transistorized Oscilloscope," *Hewlett-Packard J.*, vol. 17, no. 12, August, 1966.
34. Carlson, R., S. Krakauer, K. Magleby, R. Monnier, V. Van Duzer, and R. Woodbury: Sampling Oscillography, *IRE WESCON Conv. Record*, vol. 3, pp. 44–51, 1959.
35. Frye, G. J.: Oscilloscope Sampling Techniques, *Electronic Inds. Tele-Tech*, vol. 24, June, 1965.
36. Howard, D. L., A. I. Best, and J. M. Umphrey: The Wide-band Sampling Gate: An Analysis, Characterization, and Application Discussion, *IEEE WESCON Conv. Record, Session* 23, August, 1966.
37. Zimmerman, A.: "The State of the Art in Sampling," *Service Scope*, no. 53, Tektronix, Inc., October, 1968.
38. Grove, W. M.: "Sampling for Oscilloscopes and Other RF Systems: DC through X-band," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-14, pp. 629–635, December, 1966.
39. Frye, G. J., and N. S. Nahman: "Random Sampling Oscillography," *IEEE Trans. Instr. Meas.*, vol. IM-13, pp. 8–13, March, 1964.
40. Zimmerman, A.: "The Random Sampling Oscilloscope," Tektronix, Inc., pamphlet, 1967.
41. Luscher, R.: "Time-base Circuit for Sampling Scope," *IEEE Trans. Instr. Meas.*, vol. IM-17, pp. 111–115, June, 1968.
42. Katzmann, F. L.: "A Time-shared Oscilloscope with Precise Digital Readout," *IEEE Int. Conv. Record*, vol. 11, pp. 219–224, 1963.
43. Moffitt, Lee: Time-domain Reflectometry—Theory and Applications, *EDN Test Instr. Ref. Issue*, 1964.
44. Hewlett-Packard, Inc., Application Note 62.
45. Unter, Brian D.: Fully Calibrated Frequency-domain Measurements, *Hewlett-Packard J.*, vol. 19, August, 1968.

46. Siegfried, L.: Frequency-domain Oscilloscope Now Measures to 1,250 MHz, *Hewlett-Packard J.*, vol. 20, April, 1969.
47. Riggert, H. R.: The Role of Electronic Medical Instrumentation in Patient Monitoring, *Hewlett-Packard J.*, vol. 18, June, 1967.
48. Issacs, J. H., W. A. Peterson, and P. B. Gee: A Complete Instrumentation System for Vectorcardiography and Electrocardiography, *Dig. Int. Conf. Med. Electron. Biol. Eng.*, (6*th*) Tokyo, pp. 6–8, 1965.
49. Marinacci, A. A.: "Applied Electromyography," Lea & Febiger, Philadelphia, 1968.
50. Bryden, J.: Design Considerations for Computer Driven CRT Displays, *Computer Design*, March, 1969.
51. Johnson, A. D., and D. G. Cowden: Considerations in Specifying Display System CRT Design Objectives, *Inform. Display*, May-June, 1967.
52. Stadtfeld, N.: "Information Display Concepts," Tektronix, Inc., Beaverton, Ore., 1968.
53. House, C.: Large-screen High-frequency X-Y-Z Display, *Hewlett-Packard J.*, vol. 19, December, 1967.
54. Winningstad, C. N.: The Simplified Direct-view Bistable Storage Tube in Computer Output Applications, *Proc. Natl. Symp.* (8*th*), *Soc. Inform. Display*, pp. 129–136, May, 1967.
55. A New Look in Information Display, *Tekscope*, vol. 1, June, 1969.
56. Kurshaw, J., R. D. Lohman, and G. B. Herzog: Cathode-ray Tube Plots Transistor Curves, *Electronics*, vol. 26, pp. 122–127, February, 1953.
57. Knapton, James, and Jerrold Rogers: A New Dimension in Curve Tracing, *Tekscope*, vol. 1, February, 1969.
58. Winningstad, C. N.: In and Out of Circuits with Probes, *Proc. Natl. Electron. Conf., Chicago*, vol. 19, pp. 164–172, 1963.
59. Kohl, Wayne A.: Errors in High-frequency Oscilloscope Measurements, *Hewlett-Packard J.*, vol. 20, November, 1968.
60. Forge, Charles O.: A New Clip-on Oscilloscope/Voltmeter Probe, *Hewlett-Packard J.*, vol. 11, July, 1960.
61. Kan, George S.: New Uses for Fluxgate Principle, *Electronic Inds. Tele-Tech*, August, 1960.
62. Tyler, R. W., C. Straub, V. I. Saunders, F. C. Eisen, and T. Gentry Veal: Photography and Photometry of Cathode-ray-tube Displays, *Phot. Sci. Eng.*, vol. 7, pp. 289–304, September-October, 1963.
63. Bird, G. R., and Allan E. Ames: High-speed Oscillography with Transfer Films: A System Analysis, Paper 342, Polaroid Research Laboratories, 730 Main, Cambridge, Mass.
64. Gilbert, B.: An X-Y Plotter Stabilizing Adapter for Use with Sampling Oscilloscopes, *Mullard Tech. Communs.*, vol. 8, pp. 90–104, December, 1964.
65. Deans, J. N.: A Simple Method for Recording Fast and Low-level Waveforms, *Hewlett-Packard J.*, vol. 17, September, 1965.

CHAPTER TWELVE

RECORDERS

Arthur Miller

Consulting Engineer

Otto S. Talle, Jr.

Hewlett-Packard Company, San Diego, California

C. D. Mee†

International Business Machines Corporation

A recorder is a measuring instrument that displays a time-varying signal in a form easy to examine and reexamine, perhaps long after the original signal has ceased to exist. This seems like a fitting definition for our purpose. Of course, some devices are called recorders even though they are not measuring or analytical instruments, for instance, the familiar tape recorder used for speech and music. In this chapter, however, the treatment is limited to *measuring* instrument recorders.

Three kinds of recorders will be studied. Perhaps the type easiest to conceive is simply a meter having an indicating needle and a writing pen attached to that needle. If a strip of paper is pulled at constant velocity under the writing pen (at a 90° angle to the direction of pen motion), the

† Dr. Mee's material was reprinted with permission of the author from Magnetic Recording Techniques, a condensed article by Mee in "Encyclopaedic Dictionary of Physics," vol. 2, p. 151, Pergamon Press, New York, 1967.

moving pen will plot something resembling the time function of the signal applied to the meter. In commercial versions of this recorder, the motor that moves the writing pen is usually a D'Arsonval movement of highly special design.

We shall call this type a *galvanometric recorder*. A familiar example is the little strip-chart recorder used in electrocardiographs to plot the potentials produced at various points on the body by the contractions of the heart muscles. Some properties of the galvanometric recorder are a frequency range from zero to several kilohertz, simplicity, fair accuracy, and usually moderate cost.

Another recorder type is basically a voltage-responsive positional servo using a motor to move a writing device back and forth across a piece of paper. The servo system can be made extremely accurate, rugged, and as powerful as required. If two servos drive a single pen orthogonally over the surface of a *stationary* piece of paper, an *xy recorder* is produced. This type is capable of plotting one signal versus another in rectangular coordinates. Frequency response is limited by the maximum slewing speed of the moving system, which is about 20 in./sec in a modern *xy* recorder.

The *magnetic recorder* is the third type to be considered. In it, a thin magnetic tape, or sometimes a wire, is magnetized in accordance with a varying signal as the tape or wire passes rapidly across a magnetic recording "head." Great lengths of the tape, with recorded signals, can be wound on a reel. Later, the reel is unwound as the tape traverses a reproducing head that responds to the varying magnetic field previously recorded. Depending upon the intended application, the frequency response of magnetic recorders can extend from 0 Hz on the low end to an upper limit of from a few kilohertz to nearly 10 MHz.

Because of the wide bandwidth of tape recorders, several modes of recording—direct, FM, and digital—can be used. Magnetic recording is extremely versatile because of the bandwidth and other characteristics, such as information packing density, the capability of accepting many signals on adjacent parallel tracks, the replay of recorded signals in their original electrical form as often as desired, and the ease of erasure of the recorded signals. Unlike the other recorders, the magnetic recorder does not display the recorded information directly; the record must be processed by passing the magnetic medium across a playback transducer.

12-1 Galvanometric Recorders

Discussion in this section is limited to instruments that plot voltage or current as a function of time. Furthermore, time is represented by the uniform motion of the recording medium (usually paper) past a pointed writing device that is driven by a moving-coil galvanometer, as shown in

Fig. 12-1. The symbols and quantities to be used in the analysis of this instrument follow:

N = number of turns in galvanometer coil
l = length of long sides of coil, cm
b = breadth of coil, cm
B = flux density in air gap, gauss
e = applied signal, V
i = coil current, A
R_s = resistance of source, Ω
R_c = resistance of coil, Ω
L_c = inductance of coil, H
I = moment of inertia of coil and its mechanical load, g-cm²
τ = torque, dyn-cm
θ = deflection of coil about its axis, rad
S = torsional stiffness of coil suspension, dyn-cm/rad
R_F = mechanical resistance, dyn-cm-sec/rad

For uniform flux density in the air gap, the torque due to a given coil current and the back emf generated in the coil by its motion are both independent of coil position. The torque developed by the coil current is

$$\tau = \frac{BNlbi}{10}$$
$$= Gi \qquad (12\text{-}1\text{-}1)$$
$$G = BNlb \quad \text{abvolt-sec/rad} \qquad (12\text{-}1\text{-}2)$$

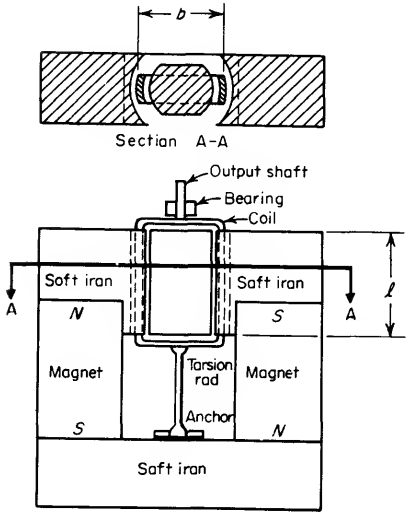


FIG 12-1 Moving-coil galvanometer.

where G is the electromagnetic coupling constant, which is characteristic of a coil of given dimensions immersed in a particular magnetic field.

The applied torque is balanced by the inertia and the frictional and stiffness torques, or

$$\frac{Gi}{10} = I\frac{d^2\theta}{dt^2} + R_F\frac{d\theta}{dt} + S\theta \tag{12-1-3}$$

For the electrical circuit,

$$e = (R_s + R_c)i + L_c\frac{di}{dt} + G\frac{d\theta}{dt}\,10^{-8} \tag{12-1-4}$$

where e is in practical volts.

In most practical galvanometers the $L_c(di/dt)$ term is small and can be neglected. In that case, we have

$$i = \frac{e}{R_s + R_c} - \frac{G \times 10^{-8}}{R_s + R_c}\frac{d\theta}{dt} \tag{12-1-5}$$

Inserting (12-1-5) into (12-1-3) gives

$$\frac{Ge}{10(R_s + R_c)} = I\frac{d^2\theta}{dt^2} + \left(R_F + \frac{G^2 \times 10^{-9}}{R_s + R_c}\right)\frac{d\theta}{dt} + S\theta \tag{12-1-6}$$

$$R_F + \frac{G^2 \times 10^{-9}}{R_s + R_c} = D \tag{12-1-7}$$

Then

$$\frac{Ge}{10(R_s + R_c)} = I\frac{d^2\theta}{dt^2} + D\frac{d\theta}{dt} + S\theta \tag{12-1-8}$$

Transient Response. If the input voltage e is a step function and if we wait long enough for the transient effects to disappear, then Eq. (12-1-8) becomes simply

$$\frac{GE}{10(R_s + R_c)} = S\theta_f \tag{12-1-9}$$

where E is the steady-state voltage and θ_f is the final deflection, or

$$\theta_f = \frac{GE}{10S(R_s + R_c)} \tag{12-1-10}$$

Equation (12-1-8) is the familiar differential equation describing the motion of a simple mechanical oscillator. Its solution shows either an oscillatory or nonoscillatory behavior depending on the magnitude of the damping term D. The details of the solution are given in standard texts on differential equations and will be omitted here.

If the damping is small enough to allow oscillatory behavior, then the solution becomes

$$\frac{\theta}{\theta_f} = 1 - \frac{T}{T_N} e^{-Dt/2I} \sin\left(\frac{2\pi t}{T} + \tan^{-1}\frac{4\pi I}{DT}\right) \tag{12-1-11}$$

Here the undamped natural period of the system is

$$T_N = 2\pi\left(\frac{I}{S}\right)^{1/2} \tag{12-1-12}$$

and the period of the actual damped oscillatory motion is

$$T = \frac{2\pi}{(S/I - D^2/4I^2)^{1/2}} \tag{12-1-13}$$

As D is increased, the damped period also increases until a value of D is reached for which T approaches infinity and all oscillatory behavior ceases. This level of D is the critical damping value

$$D_{\text{crit}} = 2(SI)^{1/2} \tag{12-1-14}$$

For this critical value of D the solution of Eq. (12-1-8) becomes

$$\frac{\theta}{\theta_f} = 1 - \left(1 + \frac{2\pi t}{T_N}\right) e^{-2\pi t/T_N} \tag{12-1-15}$$

When damping is less than critical, the response to the step input exhibits the ringing described by the sine term of Eq. (12-1-11), and the peaks and valleys occur at the times $T/2$, T, $3T/2$, and so forth. For a practical recorder the damping is usually sufficient to make all but the first peak negligible. At this time $T/2$, then, Eq. (12-1-11) reduces to

$$\frac{\theta}{\theta_f} = 1 + e^{-DT/4I} \tag{12-1-16}$$

so that the overshoot is simply $e^{-DT/4I}$.

Damping is often specified as a *percentage of critical.* This percentage can be judged from the magnitude of the overshoot that is observed in the response to a step input. For this purpose Eqs. (12-1-13) and (12-1-16) can be rewritten in terms of the factor K, where

$$K = \frac{D}{D_{\text{crit}}} \tag{12-1-17}$$

$$T = \frac{T_N}{(1 - K^2)^{1/2}} \tag{12-1-18}$$

$$\frac{\theta}{\theta_f} = 1 + e^{-\pi K/(1-K^2)^{1/2}} \tag{12-1-19}$$

The magnitude of the overshoot $e^{-\pi K/(1-K^2)^{1/2}}$ is plotted as a function of K in Fig. 12-2.

The general oscillatory solution expressed in Eq. (12-1-11), when written in terms of the damping factor K, becomes

$$\frac{\theta}{\theta_f} = 1 - \frac{1}{(1 - K^2)^{1/2}} \sin \frac{2\pi(1 - K^2)^{1/2} t}{T_N} + \tan^{-1} \frac{(1 - K^2)^{1/2}}{K} \quad (12\text{-}1\text{-}20)$$

There are differences of opinion as to the damping level that will provide optimum fidelity of reproduction of complex waveforms. In commercial recorders, D is usually set between 60 and 100 percent of the critical value. A value that provides interesting properties is 70.7 percent of critical.

In Fig. 12-3 is plotted the response to a step input for critical, 70.7 percent of critical, and 60 percent of critical damping.

It is obvious that it is the fast rise of the input step that has been distorted by the recorder. The time required for the recorder to reach the deflection θ_f can serve as a figure of merit for the instrument. This simple definition of deflection time is difficult to use practically, however, because for the critically damped response the final approach to θ_f is too gradual, and for less than critical damping the deflection overshoots before finally settling slowly to θ_f.

An arbitrary definition often given for deflection time (or *rise time*, or *response time*) is the time interval between the 10 and 90 percent points on the response to a step. From this definition and the plots of Fig. 12-3,

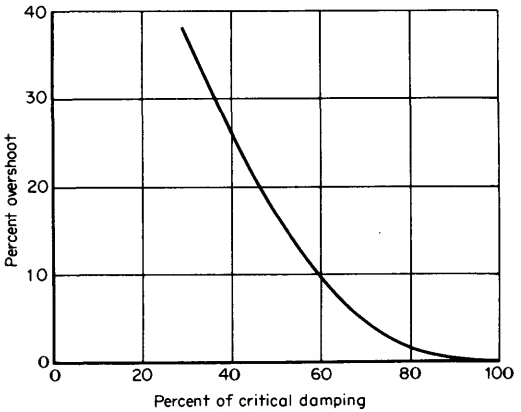


FIG 12-2 Overshoot versus percent of critical damping.

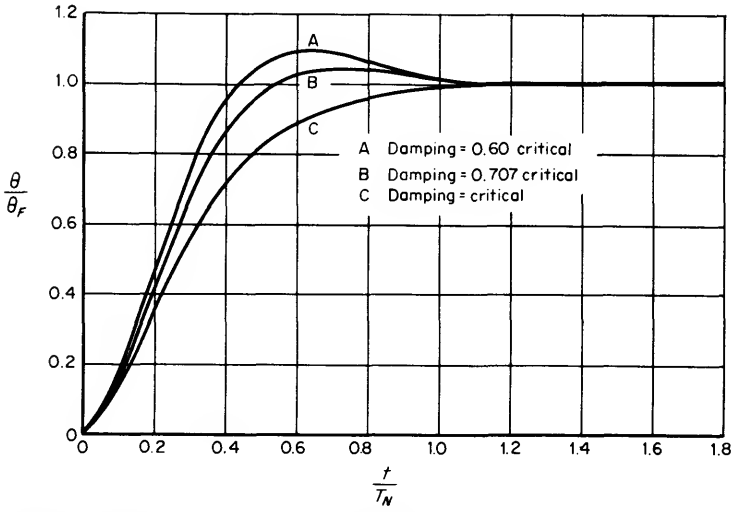


FIG 12-3 Response of galvanometer to step input.

we would have

$$T_R = 0.30T_N \quad \text{for } D = 0.60D_{\text{crit}}$$
$$T_R = 0.35T_N \quad \text{for } D = 0.707D_{\text{crit}}$$
$$T_R = 0.54T_N \quad \text{for } D = 1.00D_{\text{crit}}$$

Another test signal that is useful in evaluating the response of a galvanometer is the ramp. Since the ramp is the integral of the step, the response to the ramp can be obtained by simply integrating the response to the step. If this integration is performed for the last two damping values illustrated in Fig. 12-3 and for an input ramp that should produce a deflection of θ_f at T_n, then we get, for critical damping,

$$\frac{\theta}{\theta_f} = \frac{t}{T_N} - \frac{1}{\pi} + \left(\frac{t}{T_N} + \frac{1}{\pi}\right) e^{-2\pi t/T_N} \tag{12-1-21}$$

while, for 70.7 percent of critical damping,

$$\frac{\theta}{\theta_f} = \frac{t}{T_N} - \frac{1}{2^{1/2}\pi} + \frac{1}{2^{1/2}\pi} e^{-2^{1/2}\pi t/T_N} \cos \frac{2^{1/2}\pi t}{T_N} \tag{12-1-22}$$

These ramp responses are plotted in Fig. 12-4.

Here again the sharp corner of the input signal is distorted by the galvanometer, but after a short time interval the galvanometer does produce an undistorted copy of the input ramp, except that the reproduced

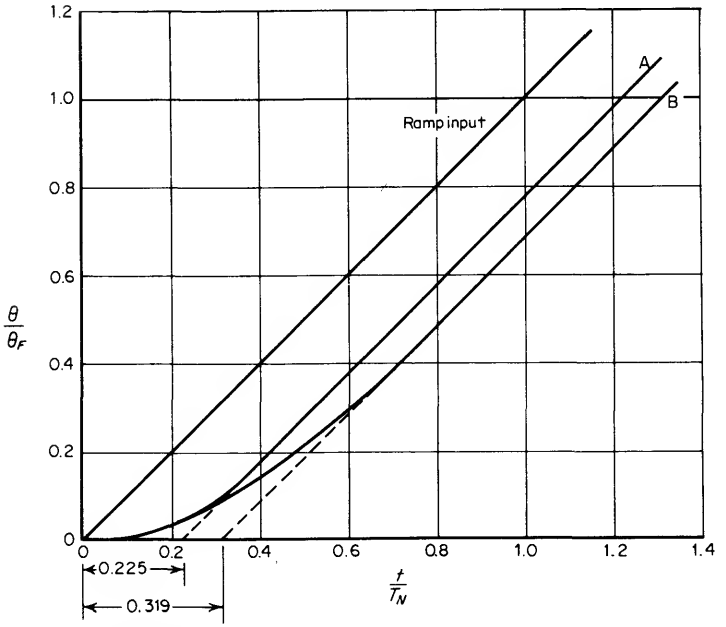


FIG 12-4 Response of galvanometer with damping equal to (A) 70.7 percent of critical value and (B) critical value.

ramp lags the input ramp by a fixed time known as the *delay time* (not to be confused with the rise time).

These delay times are obvious in Eqs. (12-1-21) and (12-1-22) and are

$$T_D = \frac{T_N}{\pi} = 0.319T_N \qquad \text{for } D = D_{\text{crit}} \tag{12-1-23}$$

$$T_D = \frac{T_N}{2^{1/2}\pi} = 0.225T_N \qquad \text{for } D = 0.707D_{\text{crit}} \tag{12-1-24}$$

The results of the ramp response can be applied to the calculation of the response to a single triangular impulse. An isosceles triangle can be considered a succession of three ramps, as shown in Fig. 12-5. From $0 < t < T_1$, the output is represented by the response to Et/T_1. From $T_1 < t < 2T_1$, the output is represented by the sum of the responses to Et/T_1 and $-2E(t - T_1)/T_1$, while for $t > 2T_1$, the output requires summing the responses to Et/T_1, $-2E(t - T_1)T_1$, and $E(t - 2T_1)/T_1$.

If we confine ourselves to triangular pulses whose base width is greater than $1.5T_N$ for the critically damped galvanometer and greater than T_N

for the slightly underdamped galvanometer, then the reproduced pulse will retain an obviously triangular shape but the pulse will appear to have been shifted bodily along the time scale by an amount T_D, the delay time.

The most obvious distortion will be the loss in amplitude of the reproduced pulse. A calculation of this loss for the critical and the 0.707-of-critical responses illustrated in Fig. 12-3 yields the results shown in the accompanying table.

Type of step response	Loss of amplitude	Time of occurrence of peak
Critically damped.....	$0.34T_N/W$	$W + 0.27T_N$
70.7% critical damping.....	$0.18T_N/W$	$W + 0.23T_N$

A typical triangle as reproduced by a 70.7 percent critically damped galvanometer is shown in Fig. 12-6. Observe that even the triangular impulse excited a very small overshoot at the trailing vertex of the tri-

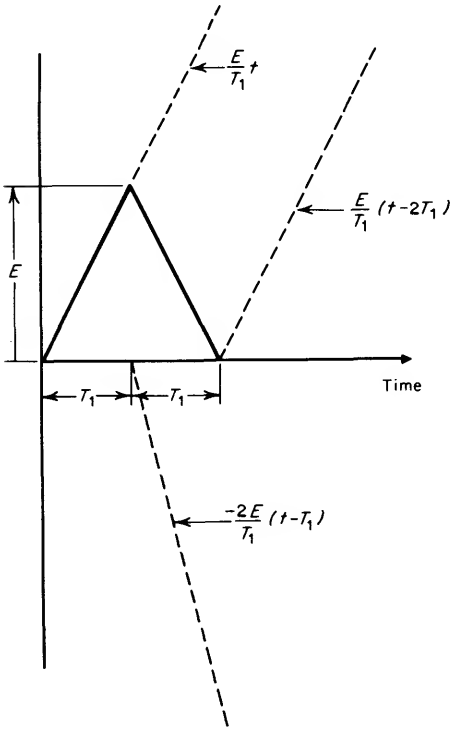


FIG 12-5 Synthesis of triangular pulse by summation of ramps.

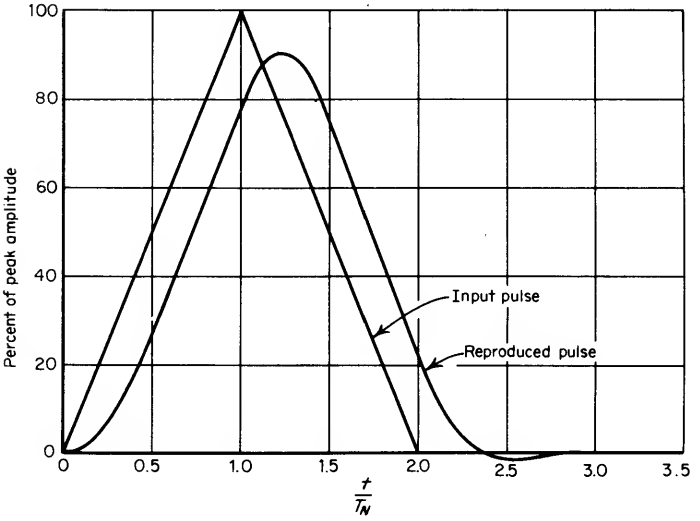


FIG 12-6 Triangular pulse reproduction, base width = $2T_N$.

angle, and that the apparent base width of the reproduced triangle has been widened.

Sinusoidal Response. The preceding analysis described the behavior of the galvanometer in the time domain for step and ramp types of input signals. Let us now consider the galvanometer response to a sinusoidal input.

The steady-state component of the solution of Eq. (12-1-8) for a sine-wave input E is

$$\theta = \frac{GE}{10S(R_s + R_c)} \frac{1}{1 - f^2/f_N^2 + j2\pi fD/S} \tag{12-1-25}$$

where θ = complex value of sinusoidal angular deflection of coil
f = applied frequency
f_N = undamped resonant frequency

At very low frequencies,

$$\theta_{lf} = \frac{G_E}{10S(R_s + R_c)} \tag{12-1-26}$$

Also, expressing D as a fraction of the critical damping value D_c, we get

$$\frac{2\pi fD}{S} = 4\pi fK\left(\frac{I}{S}\right)^{1/2} = \frac{2Kf}{f_N} \tag{12-1-27}$$

and

$$\frac{\theta}{\theta_{if}} = \frac{1}{1 - f^2/f_N^2 + j2Kf/f_N} \tag{12-1-28}$$

The magnitude of the ratio $|\theta/\theta_{if}|$ is the conventional frequency response characteristic, which is plotted in Fig. 12-7 for $K = 1.0$, 0.707, and 0.60.

Associated with the amplitude response characteristic is a phase lag

$$\gamma = \tan^{-1} \frac{2Kf/f_N}{1 - f^2/f_N^2} \tag{12-1-29}$$

This lag angle is plotted in Fig. 12-8 for the same three values of damping.

When Eq. (12-1-28) is examined as a function of the damping factor K, it is found that for small values of K, $|\theta/\theta_{if}|$ passes through a maximum greater than unity, and that as K shrinks, the magnitude of the maximum grows while its position on the frequency scale approaches f_N. Conversely, as K increases, the position of the maximum slides to the left until, at $K = \sqrt{2}/2$, the maximum occurs at $f = 0$ and the magnitude of the maximum has become unity. In other words, a damping factor of $0.707D_{crit}$ is the smallest damping for which the frequency

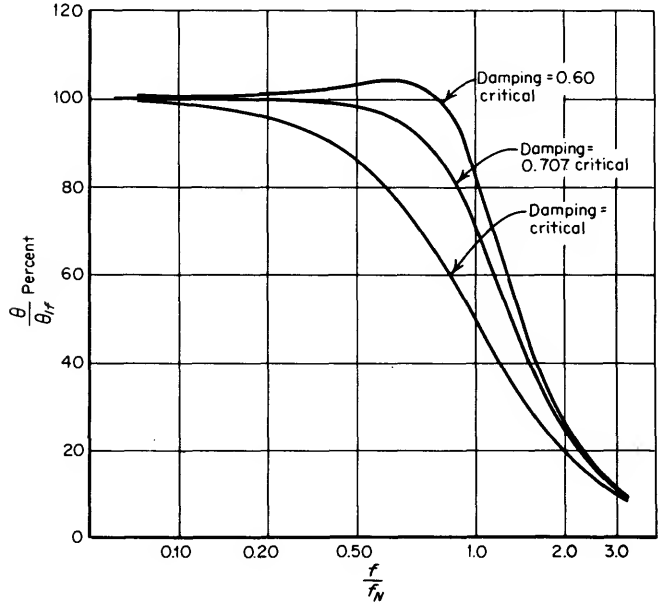


FIG 12-7 Sinusoidal response of a galvanometric recorder.

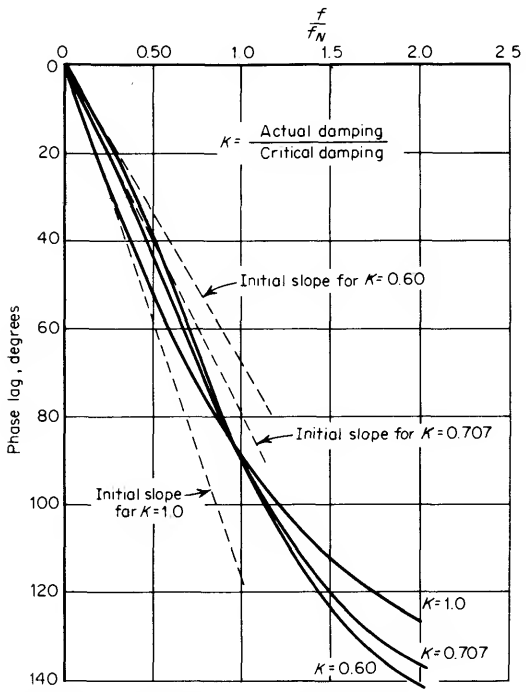


FIG 12-8 Phase lag for response in Fig. 12-7.

response curve has no bump protruding above the initial low-frequency level.

If we define the bandwidth of the galvanometer as the frequency range to the 3-dB loss point, and tabulate these values along with the corresponding rise times, we obtain the results in the accompanying table.

K	Bandwidth (BW)	Rise time	$BW \times T_R$
1.0	$0.643/f_N$	$0.54/f_N$	0.347
0.707	$1.00/f_N$	$0.35/f_N$	0.350
0.60	$1.15/f_N$	$0.30/f_N$	0.345

This table shows that over the practically useful damping range, the rise time T_R is given by the close approximation

$$T_R = \frac{0.35}{BW} \tag{12-1-30}$$

The slope of the phase lag characteristic is the delay time of the recording system.

The ideal recording system would exhibit a phase lag of constant slope and an amplitude response of constant magnitude versus frequency. From Figs. 12-7 and 12-8 it is seen that the curves for a damping factor of 0.707 show the best combination of flat amplitude response and approach to linear phase lag.

A signal that can be reproduced without serious distortion must be described by a frequency spectrum most of whose energy lies well within the passband of the recorder. The time delay associated with this portion of the spectrum is given by the *initial* slope of the curves of Fig. 12-8. This slope can be calculated from Eq. (12-1-29).

$$T_D \text{ (sec)} = \frac{1}{2\pi}\frac{d\gamma}{df} \tag{12-1-31}$$

$$\begin{aligned}\gamma &= \tan^{-1}\frac{2Kf/f_N}{1 - f^2/f_N{}^2} \\ &= \frac{2Kf/f_N}{1 - f^2/f_N{}^2} \qquad \text{for small values of } \gamma \\ &= \frac{2Kf}{f_N} \qquad \text{as } \gamma \text{ approaches zero}\end{aligned} \tag{12-1-32}$$

Thus

$$\begin{aligned}T_D &= \frac{1}{2\pi}\frac{2K}{f_N} \\ &= \frac{KT_N}{\pi}\end{aligned} \tag{12-1-33}$$

when $K = 1$,

$$T_D = \frac{T_N}{\pi}$$

and when $K = \sqrt{2}/2$,

$$T_D = \frac{\sqrt{2}\,T_N}{2\pi}$$

which are the identical values previously derived on the basis of the time-domain analysis.

Equation (12-1-28) can be solved for the value of f/f_N which will yield a response of $|\theta/\theta_{if}| = \sqrt{2}/2$ to establish the 3-dB loss points which we

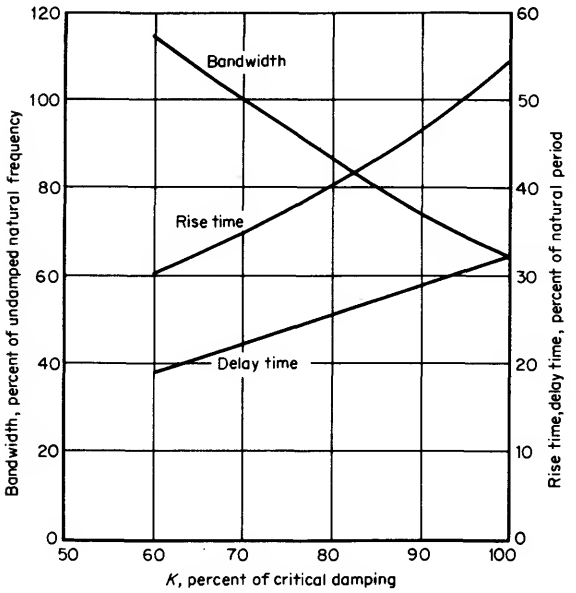


FIG 12-9 Bandwidth, rise time, and delay time as functions of damping.

have defined as the bandwidth of the galvanometer. The result is

$$\frac{f}{f_N}(-3\text{ dB}) = [1 - 2K^2 + \sqrt{4(K^4 - K^2) + 2}]^{1/2} \qquad (12\text{-}1\text{-}34)$$

From Eqs. (12-1-34), (12-1-30), and (12-1-33), we can predict the bandwidth, rise time, and delay time of a recording galvanometer, all as a function of its inherent undamped natural frequency. These relationships are summarized in the graphs in Fig. 12-9, while the overshoot in response to a step input is found in Fig. 12-2.

12-2 *Amplifiers for Galvanometric Recorders*

Consider the impedance that the galvanometer presents to the circuit that drives it. In Fig. 12-10, note that

$$E = (R_c + j\omega L_c)I_c + E_{\text{motional}} \qquad (12\text{-}2\text{-}1)$$

where currents and voltage are phasors. The last term E_{motional} is the sinusoidal back emf generated in the coil because of its motion in the

magnetic field and is

$$E_{\text{motional}} = G \times 10^{-8} \times \text{angular velocity of coil} \tag{12-2-2}$$

The angular velocity, in Eq. (12-2-2) is

$$\text{Angular velocity} = \frac{\tau}{jI\omega - jS/\omega + R_F} \tag{12-2-3}$$

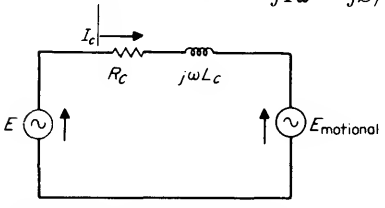


FIG 12-10 Equivalent electrical impedance of a galvanometer.

and $\tau = GI/10$, which makes

$$\begin{aligned} E &= (R_c + j\omega L)I_c + \frac{G^2 \times 10^{-9} I_c}{jI\omega - jS/\omega + R_F} \\ &= Z_c I_c + Z_{\text{motional}} I_c \end{aligned} \tag{12-2-4}$$

and (I_c is current, I is moment of inertia)

$$Z_{\text{motional}} = \frac{G^2 \times 10^{-9}}{jI\omega - jS/\omega + R_F} \tag{12-2-5}$$

The impedance of a parallel RLC circuit is

$$Z_i = \frac{1}{jC\omega - j/L\omega + 1/R} \tag{12-2-6}$$

which is identical with Eq. (12-2-5) if

$$G^2 \times 10^{-9} C\omega = I\omega$$

$$G^2 \times 10^{-9} \frac{1}{L\omega} = \frac{S}{\omega}$$

and

$$G^2 \times 10^{-9} \frac{1}{R} = R_F$$

or

$$C = \frac{I}{G^2 \times 10^{-9}}$$

$$L = \frac{G^2 \times 10^{-9}}{S}$$

and

$$R = \frac{G^2 \times 10^{-9}}{R_F}$$

The equivalent circuit of the galvanometer and drive source, therefore, appears as in Fig. 12-11. Note that the galvanometer inertia is translated into a capacitive reactance, the torsional stiffness appears inductive, and the friction, which is a series element in the mechanical system, has become a shunt element of the equivalent electrical circuit.

The voltage developed across the motional impedance represented by the parallel resonant circuit is

$$E_{\text{motional}} = jG\omega\theta \times 10^{-8} \tag{12-2-7}$$

From Eq. (12-2-7) and the circuit in Fig. 12-11 and again neglecting L_c, we can calculate the deflection θ as a function of frequency, the applied voltage E, and the galvanometer constants, obtaining

$$\theta = \frac{EG \times 10^{-1}}{(R_s + R_c)S} \frac{1}{1 - f^2/f_N{}^2 + j\dfrac{R}{2\pi f_N I}\dfrac{f}{f_N}} \tag{12-2-8}$$

where

$$R = R_F + \frac{G^2 \times 10^{-9}}{R_s + R_c} \tag{12-2-9}$$

In terms of the damping factor K, it can be shown that

$$\frac{E}{\theta} = \frac{10(R_s + R_c)S}{G}\left(1 - \frac{f^2}{f_N{}^2} + j2K\frac{f}{f_N}\right) \tag{12-2-10}$$

If this is normalized in terms of the voltage required at very low fre-

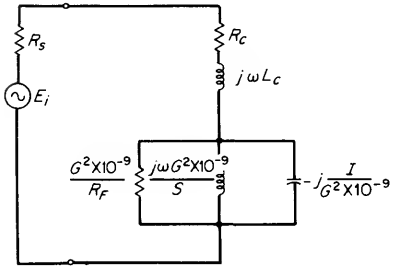


FIG 12-11 Equivalent circuit of galvanometer.

quencies, we get

$$\frac{E/\theta}{E/\theta_{lf}} = 1 - \frac{f^2}{f_N{}^2} + j2K\frac{f}{f_N} \tag{12-2-11}$$

Similarly, a calculation of the current required per radian of angular deflection gives

$$\frac{I_c/\theta}{I_c/\theta_{lf}} = 1 - \frac{f^2}{f_N{}^2} + j\frac{R_F}{2\pi f_N I}\frac{f}{f_N} \tag{12-2-12}$$

Away from the natural frequency and for small frictional constant R_F,

$$\frac{I_c/\theta}{I_c/\theta_{lf}} = 1 - \frac{f^2}{f_N{}^2} \tag{12-2-13}$$

If we are concerned with the voltage that must be delivered by a galvanometer-driver amplifier to the actual galvanometer terminals, then Eq. (12-2-11) should be calculated for $R_s = 0$

$$\frac{E_{\text{terminal}}/\theta}{E_{\text{terminal}}/\theta_{lf}} = 1 - \frac{f^2}{f_N{}^2} + j2K'\frac{f}{f_N} \tag{12-2-14}$$

where K' is the damping factor which would be obtained if the galvanometer were fed from a zero impedance source.

For galvanometers used in direct writing recorders typical values of K' are around unity.

Equations (12-2-13) and (12-2-14) are important when amplifiers are being designed that are intended to drive galvanometers. They are plotted in Fig. 12-12, the voltage plot being based on a value of unity for K'. It is seen that if full-scale deflection is to be maintained at or above the undamped natural frequency of the galvanometer, the amplifier which drives it must be capable of delivering an output voltage considerably in excess of that which could maintain the same full-scale deflection at low frequencies.

In addition to designing a galvanometer-driving amplifier for adequate output voltage capability, there is the problem of damping and gain stability to consider. The low-frequency sensitivity is established by the sum of the galvanometer coil resistance R_c and the amplifier output resistance (R_s in Fig. 12-11).

The typical galvanometer that is designed to swing a stylus for a direct writing recording process is a relatively high-power device which can and does experience a substantial temperature rise owing to its own power dissipation. This temperature rise can reach values of 40°C, and as a result, the resistance of the copper winding will increase by 16 percent over its initial value.

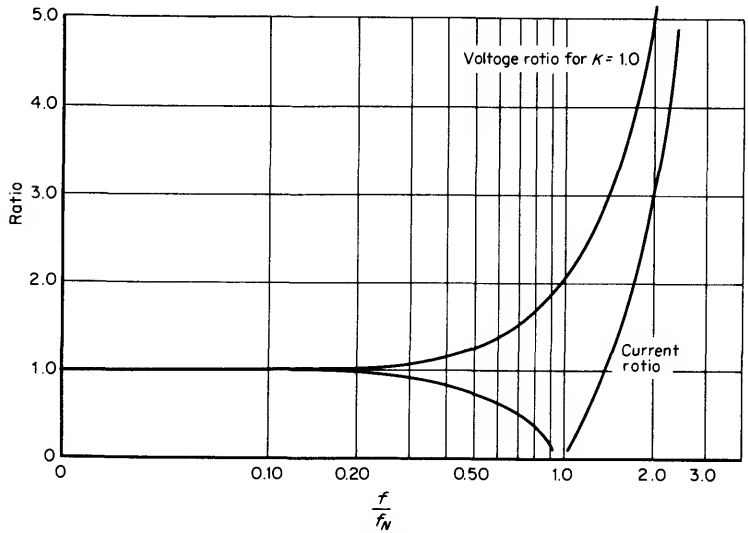


FIG 12-12 Ratios of currents and voltage for low-frequency values.

In order to make this resistance change unimportant, it is obvious that the source resistance R_s should be made very large, a requirement which can be met by the equally obvious use of current feedback around the driver amplifier. However, the typical direct-writing galvanometer requires a low value of source resistance to provide a reasonable damping factor.

These two mutually conflicting requirements can be resolved in several ways. One popular method is illustrated in Fig. 12-13. If there is high gain A in the amplifier, there will be a negligible difference between input voltage e_i and feedback voltage e_{fb}. The feedback voltage will contain two components, one of which is due to the amplifier output current i_o flowing in the feedback resistor R_{fb}, while the second component is gen-

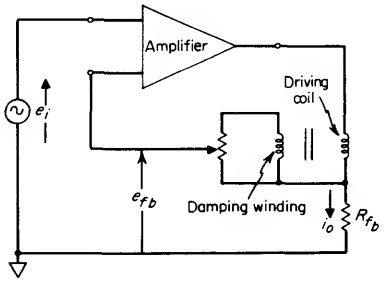


FIG 12-13 Feedback amplifier for galvanometer.

erated by the damping winding. This winding is on the same structure that carries the driving coil and is immersed in the same field, so that the voltage developed will be proportional to the galvanometer angular velocity.

$$e_i = e_{fb} = i_o R_{fb} + kG \times 10^{-8} \frac{d\theta}{dt} \tag{12-2-15}$$

or

$$i_o = \frac{e_i}{R_{fb}} - \frac{kG \times 10^{-8} d\theta/dt}{R_{fb}} \tag{12-2-16}$$

Substituting this value of i in Eq. (12-1-3) gives

$$\frac{Ge_i}{10R_{fb}} = I \frac{d^2\theta}{dt^2} + \left(R_F + \frac{kG^2 \times 10^{-9}}{R_{fb}}\right) \frac{d\theta}{dt} + S\theta \tag{12-2-17}$$

Equation (12-2-17) is identical in form with Eq. (12-1-6), but the coil resistance R_c does not appear. Instead, the sensitivity is controlled only by the feedback resistor, while the damping depends on that resistor and the setting of the potentiometer across the damping winding.

The curves of Fig. 12-12 show that if the galvanometer-driving amplifier were equalized so that, beyond the natural frequency, the galvanometer current rose rapidly with frequency, it would be possible to stretch the range of flat system response beyond the galvanometer natural frequency.

The required current rises quite rapidly, however, and soon the amplifier cannot deliver enough of either current or voltage, or the required current will burn out the galvanometer coil. Thus, if a particular galvanometer could tolerate a sinusoidal input current twice as great as that which would produce full-scale deflection at low frequencies, and the amplifier output were limited to that value, then it would be theoretically possible to equalize the system to provide flat response at full scale out to 1.75 times the natural resonant frequency of the galvanometer alone.

The equalizer could, of course, be designed to match the Fig. 12-12 requirements out to two or three times the natural resonant frequency, but, unless the signal amplitude were restricted, the amplifier output limitation would prevent the equalizer demands from being met.

Thus, it is common to find an equalized recording system specified as having flat response over a relatively wide frequency range, providing the instrument is recording a small sinusoidal signal, while the same instrument would exhibit a flat response over a much smaller frequency range if an attempt were made to record a full-scale sinusoidal signal.

A typical equalizer consists of a low-pass filter that is added to the feedback loop of the circuit of Fig. 12-13, as shown in Fig. 12-14.

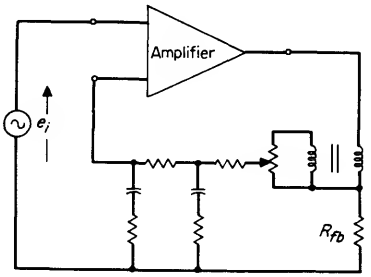


FIG 12-14 Equalized feedback amplifier.

12-3 Pen-driving Mechanisms

The conversion of the rotary motion of the galvanometer coil into a visible graphical record can be accomplished in a variety of ways ranging from a simple ink-carrying pen directly attached to the coil to a photographic system based on a mirror attached to the coil.

The simple pen system is illustrated schematically in Fig. 12-15. The locus described by the pen tip as the coil rotates is a circle of radius R. The distance moved by the pen tip from the system axis because of a coil rotation of θ is labeled y.

$$y = R \sin \theta \tag{12-3-1}$$

Using the series expansion for the sine function,

$$\begin{aligned} y &= R\left(\theta - \frac{\theta^3}{6} + \cdots\right) \\ &= R\theta\left(1 - \frac{\theta^2}{6} + \cdots\right) \end{aligned} \tag{12-3-2}$$

To a first approximation, therefore, the deflection y is directly proportional to the angle θ. The second term in the series is a measure of the departure from linearity. If θ is restricted to about $\frac{1}{4}$ rad, this term is

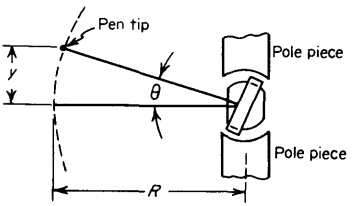


FIG 12-15 Simplest pen drive.

about 1 percent. If the system calibration is carried out at some intermediate point, say at $\theta = 0.2$ rad, then the error would be substantially reduced over the entire ± 0.25 rad assumed as full swing of the pen. This error can be completely eliminated by special printing of the chart coordinates in accordance with the sine function, but a linear scale is more attractive.

The time lines on the chart must be arcs of radius R, and in order to avoid distortion of the timing of any details drawn out by the recording pen, the galvanometer shaft must be located exactly at the center of curvature of a time-line arc. Improper positioning of the galvanometer or misalignment of the chart paper in the recorder can give the response to a step input an appearance of having either a negative rise time or an exaggeratedly long rise time, depending on the direction of the deflection.

One method of avoiding the distorted appearance of curvilinear coordinate recordings, as well as the errors in timing measurements, is to produce the recording in rectangular coordinates by utilizing the technique shown in Fig. 12-16.

In this design the chart paper is pulled over a sharp edge that defines the locus of the point of contact between the paper and the recording stylus. The stylus is rigidly attached to the galvanometer coil and wipes over the sharp edge as the coil rotates.

The paper is usually of the heat-sensitive variety, and the stylus is equipped with a heated tip long enough to guarantee a hot point of contact with the paper regardless of the stylus position on the chart. The paper could, of course, be electrically sensitive, in which case the stylus tip would serve to carry current into the paper at the point of contact.

In any case, the visible mark left on the paper has a y coordinate, measured from chart center, of

$$y = R \tan \theta \tag{12-3-3}$$

where R is the distance from the galvanometer shaft to the sharp edge,

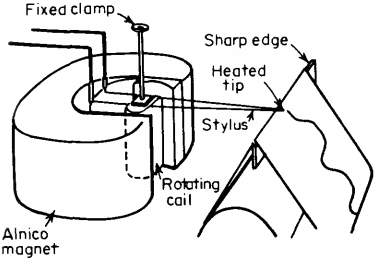


FIG 12-16 A rectangular recording arrangement.

and θ is the angle through which the galvanometer has been rotated. By using the series expansion for the tangent,

$$y = R\left(\theta + \frac{\theta^3}{3} + \cdots\right)$$
$$= R\theta\left(1 + \frac{\theta^2}{3} + \cdots\right) \tag{12-3-4}$$

Again, to a first approximation, y is proportional to θ. The departure from linearity is measured by the term $\theta^2/3$, which if θ is limited to a maximum of $\frac{1}{4}$ rad, would imply an error of 2 percent. If, however, the system calibration is carried out at an intermediate point, the trigonometric error would be substantially reduced. For a calibration deflection from 0 to 0.2 rad in a system for which full scale represents a swing of ± 0.25 rad, the maximum error would occur at about ± 0.11 rad and at the chart edges at ± 0.25 rad. The first of these errors would be 0.2 percent of full scale, while the error at chart edge would be 0.37 percent of full scale.

In calculating these deflection errors we were considering only the pen geometry, and we assumed an absolutely linear relationship between current in the galvanometer coil and resulting rotation θ. For dc or low-frequency ac currents, such linearity depends upon providing a torsion element for the galvanometer that maintains strict proportionality between applied torque and resulting angular deflection and upon maintaining perfect uniformity of magnetic field over the entire region that is traversed by the coil in its travel between the pole pieces.

In a practical galvanometer design, perfect field uniformity cannot be achieved because of flux fringing at the tips of the pole pieces. This diminution in average field strength as the coil approaches the pole tips will reduce the torque and deflection.

For the sine-law responsive system described first, this will aggravate the error, because Eq. (12-3-2) showed a negative error term. For the tangent-law system, there is a tendency for the errors to offset each other. Thus, for the tangent-law mechanism of Fig. 12-16, it is possible to achieve an overall linearity error in production instruments of one-half of 1 percent of full-scale deflection, for deflections of ± 0.25 rad.

When the D'Arsonval moving-coil galvanometer is built in miniaturized form, as it is for the usual optical recording oscillograph, the coil is so small that there is no room for any iron or pole pieces within the coil, and the construction is as sketched in Fig. 12-17. Figure 12-17*b* illustrates the relative orientations of the coil and magnetic field vector when the coil is turned through an angle α in relation to its neutral position. Because the force developed by the coil current is no longer perpendicular to the

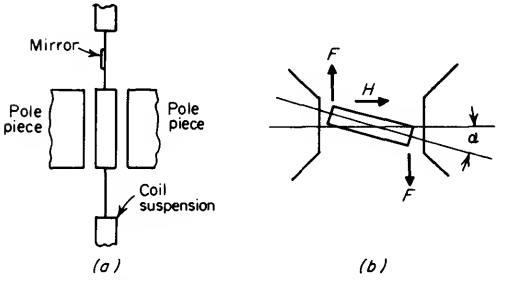


FIG 12-17 Galvanometer for optical recording: (*a*) elevation and (*b*) magnified plan view.

coil axis, the torque is

$$\tau = \frac{Gi}{10}\cos\alpha$$

$$= \frac{Gi}{10}\left(1 - \frac{\alpha^2}{2} + \cdots\right) \tag{12-3-5}$$

For dc or low-frequency signals this torque will produce a coil rotation

$$\alpha = \frac{\tau}{S} \tag{12-3-6}$$

$$= \frac{Gi}{10S}\left(1 - \frac{\alpha^2}{2} + \cdots\right) \tag{12-3-7}$$

Since the term $\alpha^2/2$ is a correction of small magnitude, it is permissible to replace it with its approximate value $G^2i^2/200S^2$, so that

$$\alpha = \frac{Gi}{10S}\left(1 - \frac{G^2i^2}{200S^2} + \cdots\right) \tag{12-3-8}$$

A light beam hits the galvanometer mirror and is reflected from the mirror onto the recording medium. As the mirror rotates through the angle α, the reflected beam will swing through the angle θ. If the distance from the mirror to the record is R, then the recorded deflection will be

$$y = R\tan\theta$$

$$= R\tan 2\alpha$$

$$= R\left(2\alpha + \frac{8\alpha^3}{3} + \cdots\right)$$

$$= 2R\left(\alpha + \frac{4\alpha^3}{3} + \cdots\right) \tag{12-3-9}$$

Again, since the α^3 term is itself only a correction, we can approximate it well enough with $G^3i^3/10^3S^3$. Thus,

$$y = 2R\left[\frac{Gi}{10S}\left(1 - \frac{G^2i^2}{200S^2}\right) + \frac{4(G^3i^3)}{3 \times 10^3S^3}\right]$$
$$= \frac{2RGi}{10S}\left(1 + \frac{5}{6}\frac{G^2i^2}{100S^2}\right) \qquad (12\text{-}3\text{-}10)$$

In terms of the angle θ, the last expression becomes

$$y = \frac{2RGi}{10S}\ (1 + \tfrac{5}{24}\theta^2) \qquad (12\text{-}3\text{-}11)$$

Comparing this with the performance of the direct-writing tangent-law recorder characterized by Eq. (12-3-4) shows how the excess deflection resulting from a strict adherence to the tangent law has been reduced by the compensatory effect of the torque's falling off in accordance with the cosine function.

The small galvanometer used in optical recorders differs from the large units found in direct writers in another important respect, namely, the method of achieving proper damping. From Eqs. (12-1-7) and (12-1-14), the condition of critical damping is given by

$$R_F + \frac{G^2 \times 10^{-9}}{R_s + R_c} = 2\sqrt{SI} \qquad (12\text{-}3\text{-}12)$$

If the mechanical friction is negligibly small, then the damping must come from the term $(G^2 \times 10^{-9})/(R_s + R_c)$. When, however, the stiffness S is raised to the values required to obtain a high natural frequency, then the required corresponding source resistance R_s becomes either zero or negative, or the value of G becomes so high that the necessary flux density cannot be reached in a practical structure.

For these reasons, optical oscillograph galvanometers usually depend on magnetic damping in units whose natural frequency is 500 Hz or less. When the natural frequency exceeds 500 Hz, the required damping is derived from the mechanical friction term R_F. This friction is obtained by immersing the moving coil in a viscous oil. The viscosity is chosen to provide a damping factor somewhere between 60 and 70 percent of critical. As the natural frequency of the galvanometer is increased, the overall damping becomes more and more dependent on the oil and less and less on the circuit resistance.

Linkages for Straight-line Writing. Although the method of obtaining direct-writing recording in rectangular coordinates shown in Fig. 12-16 is simple, it can be applied easily only to heat-sensitive or electrically sensitive recording papers. When the recording is to be made by ink, the

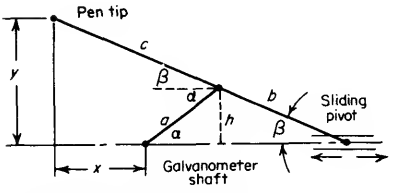


FIG 12-18 Straight-line pen mechanism.

pen tip must usually be associated with some kind of mechanical linkage that can convert the rotary motion of the galvanometer shaft into a straight-line motion of the pen tip and that will maintain some reasonable proportionality between the pen tip displacement and the galvanometer angular displacement.

A typical linkage of this sort is shown in Fig. 12-18. In this linkage the galvanometer shaft carries a crank a, which is attached to the pen at a pivot point. The pen length from this pivot to the writing tip is c, and the length to its rear end is b. This rear end is constrained to slide along a straight line which passes through the galvanometer shaft. This straight line establishes the axis of the linkage.

Assuming small values of α so that only the first two terms of the cosine expansion are significant, it can be proved that x is constant as α and y vary if

$$c = \frac{b^2}{a} \tag{12-3-13}$$

Actually, the motion of the pen does not stay perpendicular to the direction of motion of the recording paper because the above assumption becomes worse as α increases. A value of c slightly greater than that shown in Eq. (12-3-13) is a better compromise. The nonlinearity

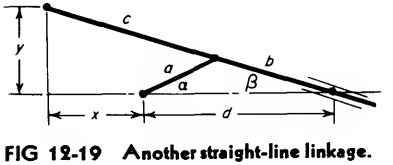


FIG 12-19 Another straight-line linkage.

between y and deflection angle in Fig. 12-18 is the same as that in Fig. 12-15 if we let the pen length $R = a(1 + c/b)$.

Another straight-line linkage is shown in Fig. 12-19. Here the rear end of the pen slides through a guide which is free to rotate about a

fixed axis. The galvanometer may be arranged to drive the crank a, or the galvanometer may drive the pen guide.

Here again it turns out that the required length of the link is nominally b^2/a, but that a slight increase in the length of c beyond this nominal value is desirable. The distance from the rotating guide to the writing line is $c - a + d$. The deflection is then

$$\begin{aligned} y &= (c - a + d) \tan \theta \\ &= (c - a + d) \tan \sin^{-1} \frac{a \sin \theta}{(d^2 + a^2 - 2da \cos \theta)^{1/2}} \end{aligned} \tag{12-3-14}$$

In all the descriptions of recorder behavior so far we have tacitly assumed that the mechanical friction R_F was of such a nature that the braking torque it exerted on the galvanometer motion was proportional to the galvanometer's angular velocity. Unfortunately, real physical mechanisms involving shafts and bearings, or pen tips rubbing on paper surfaces, do *not* exhibit pure viscous friction. Instead, it is found that for a real mechanism, some substantial current must be applied to the galvanometer in order to get any motion at all. Also, once the galvanometer has been deflected by an input current, the removal of that current does not guarantee that the galvanometer will return to its initial position.

These frictional forces, which exist even under conditions of zero velocity, represent a form of hysteresis that can be tested by applying a deflecting current first in one direction and then in the other, with periods of zero current between the successive reversals, as illustrated in Fig. 12-20 in exaggerated form. For this test the applied signal is passed through a low-pass filter to ensure a gradual application and removal of the signal, or the galvanometer damping is set for at least critical so that there will be no overshoot. The offset δ should approach zero. In practical direct writers it can be kept less than 0.5 percent of the full chart width.

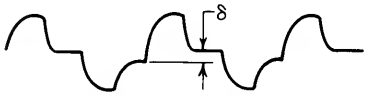


FIG 12-20 Effect of static friction in pen motors.

Another manifestation of nonlinear mechanical frictional effects is an apparent change in the waveform of signals whose amplitude is gradually reduced. Thus, triangular pulses that look like triangles when their amplitude is 1 cm on the chart might look more like sine waves when

their amplitude is reduced by a factor of 5, and when their amplitude is reduced by a factor of 10, the recorded amplitude may be substantially less than 1 mm.

The major sources of nonlinear mechanical resistance in direct-writing recorders are the friction of the pen or stylus against the paper, bearing frictions, internal frictions in the material out of which the torsion element is made, and hysteresis effects at the clamping points where the torsion element is attached to the moving coil and at the attachment to the fixed supports. Torsion bar design is a mechanical engineering problem beyond the scope of this treatise, and the same can be said of the choice of bearings.

Friction at contact of the stylus with the paper depends on the pressure of the stylus against the paper and the lubrication, if any, that exists at the contact. For example, when the recording is made by a hot stylus on the most commonly used heat-sensitive paper, the hot stylus melts the plastic coating on the paper and the molten plastic acts as a lubricant. If the stylus is insufficiently heated, this lubricating process is lost and performance is adversely affected, especially in the frequency range near the natural frequency of the galvanometer, where the stiffness and inertia mechanical reactances cancel each other and the behavior of the system is determined almost entirely by the mechanical resistance.

The quality of the written record is improved by increasing the contact pressure of the pen or stylus against the paper. This increases the friction. Recording by ink, in rectangular coordinates, requires linkages that involve additional rotating or sliding bearing contacts beyond those which support the galvanometer shaft itself. These increase the friction.

In order to make these increased frictions tolerable, one can increase the torsional stiffness of the galvanometer suspension. This leads to a higher natural frequency, which is desirable, and to reduced sensitivity, which is undesirable. In fact, the increase in required driving current may soon exceed the power dissipation capability of the galvanometer structure. An increase in apparent stiffness for small deflections without a corresponding increase in required current for large deflections can be achieved by equipping the galvanometer-amplifier combination with some kind of position feedback system. The basic idea is shown here in Fig. 12-21.

For the following simplified calculations of stiffness, let us assume a direct proportionality between stylus deflection, transducer output, and angular rotation of the galvanometer coil.

$$\text{Transducer output} = e_{fb} = K_T\theta \tag{12-3-15}$$

where K_T is the transducer sensitivity in volts per radian.

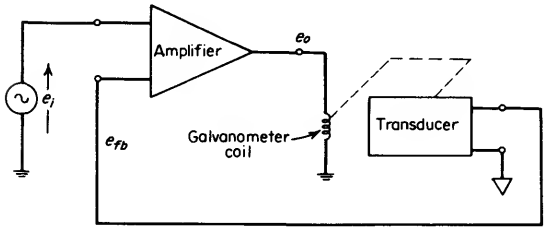


FIG 12-21 Position feedback in pen drives.

If the stylus is mechanically pushed from its resting position so that the galvanometer coil is rotated through this angle θ, then the transducer output is the net input to the amplifier. This input produces an amplifier output voltage

$$e_o = K_T A \theta \tag{12-3-16}$$

where A = amplifier voltage gain, and a corresponding galvanometer coil current

$$i = \frac{K_T \theta A}{R_c} \tag{12-3-17}$$

which generates a torque

$$\tau = \frac{Gi}{10} = \frac{GK_T \theta A}{10R_c} \tag{12-3-18}$$

Since the transducer output is phased to produce a degenerative feedback signal, this torque will counteract the initially applied push on the stylus. The apparent stiffness, then, is

$$S_{\text{apparent}} = \frac{\tau}{\theta} = \frac{GK_T A}{10R_c}$$

For example, a conventional torsion-controlled galvanometer might have a stiffness of 2×10^6 dyne-cm/rad, $G = 10^7$, and $R_c = 10$. Replacing the torsion rod by position feedback with $A = 100$, and $K_T = 10$ would give an apparent stiffness of 10^8 dyne-cm/rad, an increase of 50 times with a corresponding improvement in overcoming friction.

In a typical practical system, the stylus might be deflected 1 cm for a coil rotation of 0.1 rad, and full scale might be ± 2.5 cm (or ± 0.25 rad). These figures correspond to a transducer output of ± 2.5 V and a required input swing of the same magnitude. Thus, a discrepancy of 1 mm between the actual stylus position and that demanded by the input signal would call forth an error signal of 0.1 V, and an amplifier output of

10 V at 1 A. However, a discrepancy of 2 mm probably would not succeed in developing a correction current of 2 A because the amplifier would probably have reached its maximum output capability long before the output reached the 2-A level.

The feedback system, therefore, if it has sufficient gain to be effective, is also bound to exhibit nonlinear properties over much of the range. However, these nonlinearities may appear in ranges where they are unimportant, as in the stiffness, where we are most anxious to achieve high stiffness for small deflections, but the fact that large deflection errors are not accompanied by proportional increases in restoring torques does not matter.

If the apparent stiffness brought by position feedback is so great, then it is obvious that a conventional torsion element is not a necessary part of the galvanometer structure. This means that a given static deflection of the stylus can be maintained at any position of the chart with no steady current supplied to the galvanometer by the amplifier. In fact, the amplifier is called upon to supply power to the galvanometer only when the instantaneous stylus position does not correspond exactly with the instantaneous input signal, and this lack of correspondence will exist only when the stylus is moving.

When the torsion element is present, the galvanometer will of course exhibit a mechanical resonance, and the current required will go through a minimum at this resonance. Beyond the resonance frequency, the required current remains below that which would need to be supplied to the same galvanometer without a torsion element.

For a given maximum permissible current, therefore, a recorder with a mechanically resonant galvanometer will be capable of providing a greater bandwidth. The presence of the torsion element, however, means that some current is necessary to maintain a fixed deflection, and thus one of the advantages of the position feedback process is lost. Furthermore, if the torsional constant is large, so much current might be called for to place the stylus near the chart edge that there would not be much reserve left to make any corrections of errors sensed by the feedback transducer.

A compromise can be made which would require substantially less than maximum permissible current to reach chart edge, but which would still provide a significant increase in bandwidth. The curve shown in Fig. 12-22 for $S = 2 \times 10^6$, for example, represents a practical limit to this compromise, and it results in an increase from 64 to 79 Hz in the maximum frequency at which a full-scale sinusoidal deflection can be maintained by a peak current of 1 A.

Although the simple schematic of Fig. 12-21 illustrates the basic idea of a position feedback recorder, it would not work as a practical matter

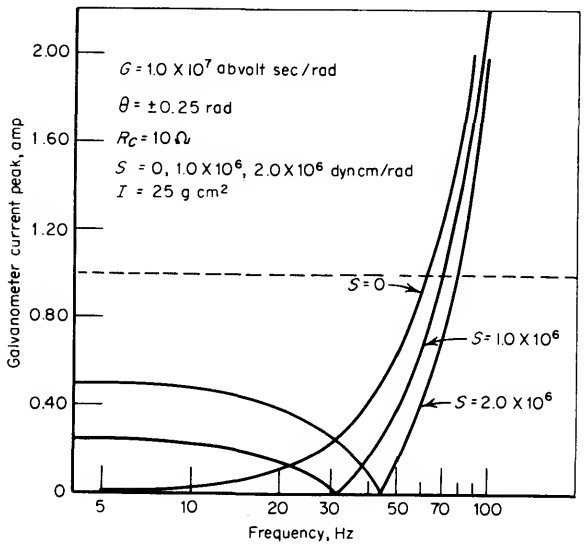


FIG 12-22 Current requirements for a galvanometer.

because the dynamic properties of the galvanometer would introduce phase shift between the amplifier output and the feedback signal developed by the transducer and such a phase shift would result in system oscillation. Conventional stabilizing procedures are employed, with proper regard for the nonlinear character of the system.

12-4 Servorecorders

Servorecorder uses fall into either one of two general types. The first type is the plotting of a variable versus time. The recorder performing this task is generally called a *strip-chart recorder*, the paper being perhaps 100 ft long (time axis) and narrow in width. The paper moves through the recorder at a uniform rate (distance per time) as the variable is being recorded. The second type is plotting one variable versus a second variable, a function performed by an *xy* recorder. By introducing a ramp voltage on one axis, the equivalent of a strip-chart recorder is obtained for limited chart lengths. The conventional paper sizes are $8\frac{1}{2} \times 11$ in. and 11×17 in. (See Fig. 12-23.)

Types of Servos. The servo systems for these recorders can take many forms. These forms vary from open- and closed-loop galvanoservos (discussed earlier in this chapter) to closed-loop null-balancing servos.

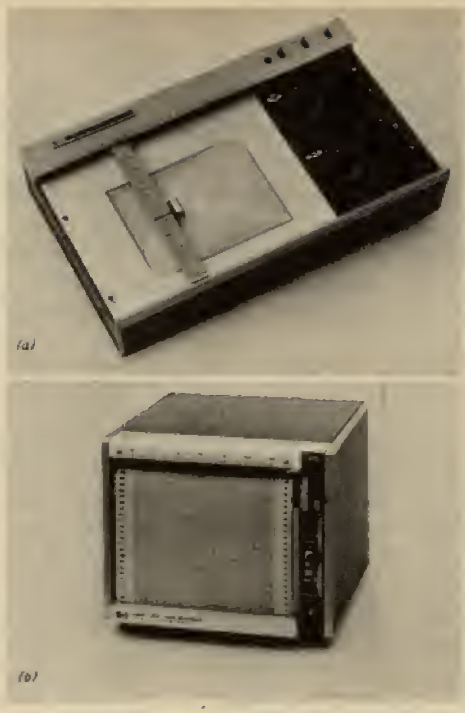


FIG 12-23 Typical null-balance servorecorders: (*a*) Hewlett-Packard Model 7035B *xy* recorder, (*b*) Hewlett-Packard Model 680M strip-chart recorder.

The most common for *xy* and strip-chart recorders are the latter, the closed-loop null-balance servo with a servomotor as the pen driver. It

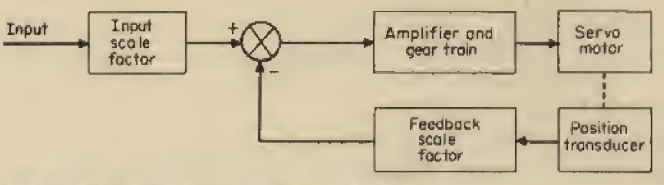


FIG 12-24 Block diagram of a null-balance servo.

should be noted that in a block diagram (Fig. 12-24) all the closed-loop null-balance systems are virtually identical; only the transfer functions indicate the approach and type of hardware used.

The classic type of input and balance circuit for these types of recorders

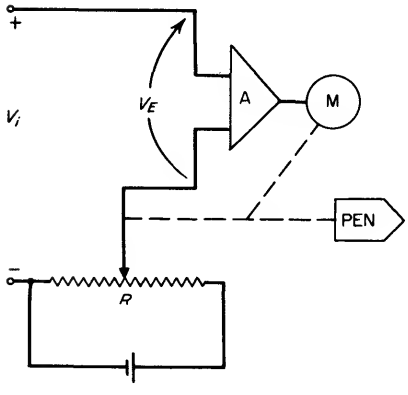


FIG 12-25 Potentiometric-input recorder.

is the potentiometric-input type, which is a closed-loop, null-balance servo system. It derives both its name and circuit (Fig. 12-25) from the laboratory dc instrument called the *potentiometer* and has four major virtues that have maintained its widespread usage. These virtues are simplicity, accuracy, stability, and zero input current at null. The servo balanced version of the potentiometer retains these characteristics. Other types of input circuits are also used, but their circuitry is often dictated by special requirements such as a need for constant input impedance, etc.

Performance Characteristics. Theoretical analysis and design of servo systems are covered in detail elsewhere in the literature and will not be directly dealt with here. However, there are some unique design aspects that must be considered when designing a servo null-balanced graphic recorder that are learned by experience, and these aspects are covered here.

Much of the recorded data might be described as virtually static. Such data might be obtained while recording the temperature of a body with a large thermal time constant. Recorder characteristics that affect its performance under these static conditions are (1) accuracy, (2) linearity, and (3) resettability.

Accuracy is partially determined by the range-to-range accuracy of the input attenuator if one is used. It is also affected by the precision of the initial calibration adjustment and the precision of the markings on the graph paper.

Linearity is usually defined as terminal based. This definition, as opposed to best straight line, offers the most meaningful and useful specification. The recorder performance in linearity is largely controlled by the slidewire linearity and the loading of the slidewire potentiometer.

The element linearity is mainly a mechanical consideration determined by the manufacture of the component. Slidewire loading, however, is entirely a design constraint that can be compensated for. Linearity can be maximized for a given loading by inserting an end resistance (equal to $0.5R$) in series with the slidewire resistance element [1]. See Fig. 12-25.

Resettability can be defined as the multidirection repeatability. Parameters controlling resettability are:

1. *Deadband.* Deadband is controlled by the system loop gain. It is the amount of error signal required to produce the starting voltage on the servomotor. The starting voltage is influenced greatly by the frictional load on the motor.

2. *Backlash.* Backlash is the mechanical hysteresis between the slidewire wiper (feedback point) and the pen tip. It is therefore desirable to have the two as close together as practical.

3. *Noise.* Noise in the servo system, usually generated in the front end of the amplifier, results in a measurable disturbance at the pen tip and therefore degrades the resettability. This noise often originates with the input signal and is directly amplified with the error voltage. An input filter is frequently a necessity.

Null-balance recorders that excel in static performance are relatively easy to design. Designing a servorecorder that behaves well under dynamic conditions, however, is not as easy.

There are two prime characteristics that govern the dynamic performance of a recorder. These are:

1. *Slewing Speed.* Slewing speed along a particular axis is defined as the maximum attainable pen speed along that axis. Under dynamic conditions the servo can reach a velocity limit (slewing speed) and be unable to follow the input signal. Under such a condition the pen will lag the input and produce an error. For a sine-wave input the velocity limit is reached when the peak amplitude is

$$A > \frac{V_s}{2\pi f} \tag{12-4-1}$$

where V_s is the maximum slewing speed. Slewing speed is controlled by the type of motor (two-phase ac, dc, and so forth), its drive conditions, and the gear train between the motor and the pen. It should be noted that servos using ac servomotors have their maximum speed controlled by the power-line frequency, whereas dc servos do not.

2. *Acceleration.* Acceleration mentioned in recorder specifications is usually defined as peak acceleration. Under dynamic conditions the servo can reach acceleration limits beyond which it will be unable to respond to the rate of change of the input. For a sine-wave input the

acceleration limit is reached when the recorded amplitude is

$$A = \frac{a_0 V_s}{2\pi f(2\pi f V_s + a_0)} \quad \text{in. peak-to-peak} \tag{12-4-2}$$

where a_0 is the peak acceleration and V_s is the slewing speed.

The system inertia is the primary limit for the peak acceleration. However, even though the mass near the pen may be large, the typical gear train between the motor and pen causes the motor rotor inertia to be the major or dominant factor.

When either the velocity or acceleration limit is reached, the servo goes into nonlinear or saturated operation. A properly designed system for general use has the velocity limit and the acceleration limit compatible so that neither one singly dominates the performance limits. Since these limits are dimensionally dependent, the conventional method of using the 3-dB point to reflect dynamic performance is generally unsatisfactory.

Figure 12-26 shows the velocity-limit and acceleration-limit curves and the resulting areas of linear and nonlinear operation. These curves, of course, only hold true when the inputs are sinusoidal.

One aspect of null-balanced servos that is often overlooked both in design and application is the inherent lack of input-noise immunity. A null-balanced system depends on its amplifier's having high gain to minimize the error signal. Any input noise that appears on the error signal is also amplified. With a "tight" high-gain servo, it is possible for even 1 percent ripple on the input dc signal to saturate the amplifier completely. A common cure is to incorporate a low-pass filter in the balance circuit, but care must be taken to see that filter response does not dominate the system.

Reliability. Reliability, or lack of reliability, of certain components is often neglected. Many designers use some components with an attitude that they are the best available, even though they are unreliable. A typical example is the resistance slidewire often used as the position

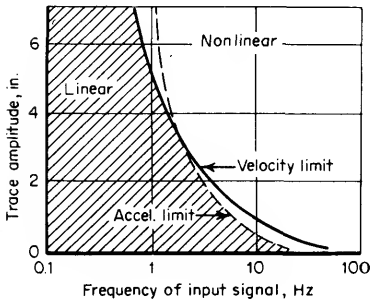


FIG 12-26 Velocity and acceleration limits.

transducer. The use of a slidewire can result in a very simple, accurate, and stable circuit. Only its reliability or life can be challenged. Two areas can cause problems: (1) mechanical wear, and (2) excessive contact resistance.

Mechanical wear is basically a design problem. If one controls the contacting or rubbing metals and the contact pressure, the life can be accurately predicted and through proper design the life can be extended.

Contact resistance that develops between the resistance element and the sliding contact is by far the most troublesome in most applications. A chemical film develops on the surface of the metals and causes the contact resistance to vary by many decades from spot to spot. This jump can be sufficient to cause spots where the servo cannot balance or even to cause the circuit to become inoperative. Two general approaches can be used to minimize the problem. A chemical film can be placed over the contacting elements to inhibit the formation of contact resistance, or the circuit can be designed to minimize the effect of the increase in the contact resistance.

Development of the proper chemical coating is no small task and depends on the metals and atmospheres being considered. The easiest and most practical method is to design the circuit so that it is immune to the development of contact resistance. Designing the circuit so that the slidewire wiper drives a minimal load, such as the gate of an FET, produces very satisfactory results.

Although it is out of the servo loop, the output medium (graph paper) should be considered. It should be noted that even though the servo is stable and accurate, if paper is used, dimensional changes of ± 1 percent can result from ambient temperature and humidity extremes. This calls for frequent recalibration or the use of a more stable base such as mylar or acetate film.

12-5 *Magnetic Recording*

Figure 12-27 shows the basic elements of a simple magnetic recording system. The magnetic tape is made of a thin sheet of tough, dimensionally stable plastic, one side of which is coated with a magnetic material. Typically, some form of finely powdered iron oxide is cemented to the plastic tape with a suitable binder. As the tape is transferred from one reel to another, it passes across a magnetizing head that impresses a residual magnetic pattern upon it in response to an amplified input signal. The signal on an exposed tape can be retrieved and played out at any time by pulling the tape across another magnetic head in which a voltage is induced.

It is possible to magnetize the tape longitudinally or along either one

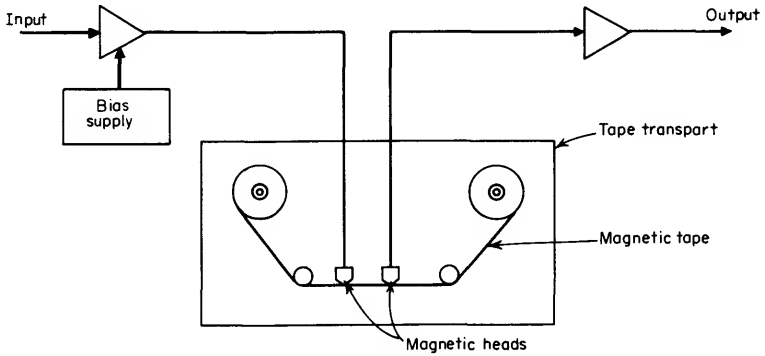


FIG 12-27 Elementary magnetic-tape recorder.

of the other two main axes, but longitudinal magnetization seems to be the best choice. This mode is assumed at present. Figure 12-28 shows in a simple way how the tape is magnetized.

Magnetic wire was used at one time, but tape has superseded the wire.

If a magnetic field is applied to any one of the iron oxide particles in a tape and then removed, a residual flux remains. The relationship between the residual flux and the recording field is determined by the previous state of magnetization and by the magnetization curves of the particular magnetic recording medium.

A single magnetic particle on the tape might have the BH relationships shown in Fig. 12-29, where H is the magnetizing force and B is the flux density in the particle. Complete degaussing of the material gives the condition at point O. Now, if the current in the coil of the recording head (Fig. 12-28) is increased from zero in a direction that gives positive

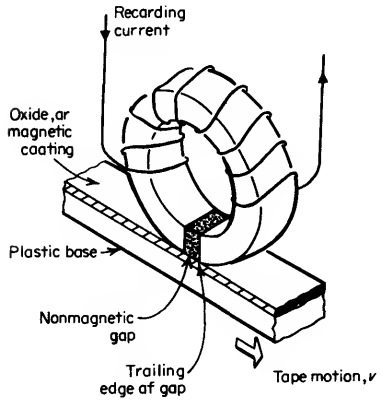


FIG 12-28 Simplified block diagram of magnetic-recording process.

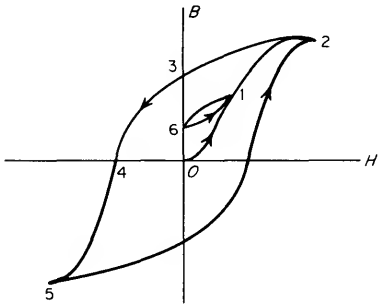


FIG 12-29 Typical magnetization curve.

values of H, the flux density increases along the path O, 1, 2, until eventually the material is saturated. If the operating point is brought from O only as far as 1, and H is then brought back to zero, B follows a minor hysteresis loop back to the point 6. A greater value of coil current would leave a higher residual flux, and a lower current a lower residual, and a very simple recording process results. However, the linearity between residual flux and recording current is extremely poor.

Below, in Sec. 12-6, a method for securing linearity in direct recording will be described. Also, the use of FM in recording, to yield independence from amplitude variations and to extend the lower frequency range to dc, will be discussed. Finally, various pulse recording schemes for instrumentation will be described.

In all systems, the signal is reproduced by passing the magnetized tape over a magnetic head similar to the recording head. The magnetization of the particles on the tape induces a varying flux in the reproducing head, and a voltage is induced in the coil, proportional to the rate of change of flux.

Uses and History of Magnetic Recording. The first patent for a magnetic recorder was issued to Valdemar Poulson in Denmark before the turn of the century. However, there was little further development or usage until about 30 years ago because of distortion problems and lack of adequate electronic circuitry to go with the magnetic structures. Between 1940 and 1950, both development and application accelerated rapidly for use in voice and music recording, instrumentation, and data processing.

The rapid acceleration was due in part to several important technical advances. One was the discovery that a high-frequency bias superimposed on the signal to be recorded could reduce distortion to the low levels required in high-quality broadcasting. Another advance was the use of an FM carrier and FM demodulation to secure low distortion, flat frequency response, and compensation for slight speed variations in tape speed. Also, magnetic recording began to be used in computer systems for memory.

All the recorders studied in this chapter are important instruments. The following inherent capabilities of *magnetic* recorders should be compared with the qualities of other types:

1. Magnetic recorders have by far the widest frequency range, that is, from dc to several megacycles.
2. Present magnetic recordings have low distortion, whether used for pulse recording, FM systems, or direct recording with a high-frequency bias.
3. The dynamic range can exceed 50 dB, and recovery from saturation is virtually instantaneous.
4. In other recorders, information is stored in *visual* form, whereas in magnetic recorders the signal is available immediately in its initial electrical form.
5. The recording medium, the tape, can be erased and used many times. Further, the tape can be spliced and edited.
6. A piece of tape can be played back repetitively to permit extensive analysis.
7. Several or many channels can be recorded simultaneously on one tape.
8. The time base of the signal can be changed by merely playing the tape back at a speed different from that for recording.
9. Magnetic recording can give exceedingly high density of data points to simplify storage and handling.

The unique features of magnetic recording suggest applications for it. Recorders are made in many sizes and are available in portable, semiportable, and rack-mounted form. A few of the applications follow:

1. Industrial research and production monitoring and control, including stress and vibration recording, fuel-consumption logging, and noise analysis
2. Data recording and analysis on aircraft, missiles, and satellites and in general flight testing
3. Communications surveillance and spying
4. Medical research and patient monitoring

Memory and programming instructions for computation.

12-6 Magnetic Recording Techniques*

The ever-increasing need for data storage systems with large capacity and fast access to the stored information is being satisfied at present by

* From Magnetic Recording Techniques, a condensed article by C. D. Mee in "Encyclopaedic Dictionary of Physics," vol. 2, p. 151, Pergamon Press, New York, 1967, with permission from the author.

a variety of magnetic recording systems. The existing magnetic recording technology combines storage stability with high storage density and fast information reversibility. Basically, all of the existing recording systems with magnetic tapes, strips, disks, drums, or loops use the same recording and reproducing transducers and magnetic medium. For these components, the techniques of recording and reproducing will be described in this article.

The Recording Process [2 to 6]. The recording process to be examined is the mechanism by which an electromagnetic recording head magnetizes the tape. The magnetic field acting on an element of tape as it passes the recording gap region varies in direction and magnitude with the position of the element as shown in Fig. 12-30. On the left-hand side of the ordinate H, the total field amplitude is plotted for different relative distances y/g from the recording-head surface, where g is the gap length. It is an objective of high-resolution recording to produce a very rapid decrement of the recording field as the tape moves away from the gap region in the x direction. Since this condition occurs near the head surface as shown in Fig. 12-30, only a thin magnetic layer can be effective for short-wave length recording. The change in direction of the recording field is illustrated by plotting the longitudinal and perpendicular components of the field on the right-hand side of the ordinate in Fig. 12-30. It can be seen that, at the gap center plane, the field is entirely longitudinal and rotates toward the perpendicular direction as a tape element moves away from the center plane.

In order to understand how the tape magnetization depends on the recording field direction, reference is made to typical hysteresis loops for a conventional oxide-powder tape. Figure 12-31*a* shows a family of loops for a field applied along the direction of orientation of the long axes of the elongated particles; this corresponds to the longitudinal head field direction in conventional tapes. It can be seen that, for small, applied fields, a low remanent magnetization is acquired. This rapidly

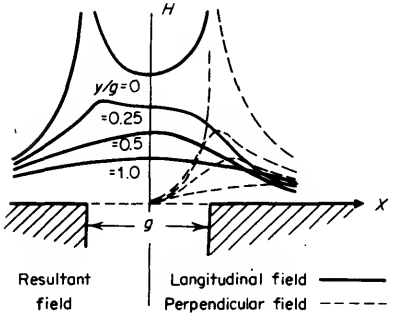


FIG 12-30 Field distribution near a recording head gap: x = direction of tape motion, y = perpendicular distance from gap, and g = gap length.

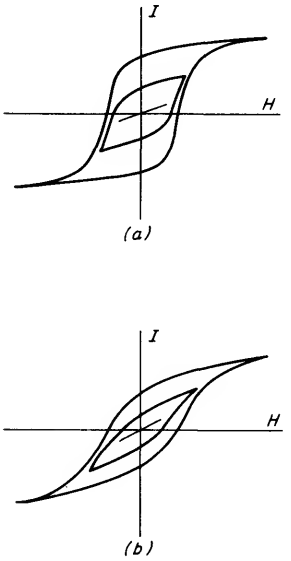


FIG 12-31 Hysteresis loops of 60 Hz for oriented particles of γ-Fe_2O_3; H_{max} = 1,000 Oe: (*a*) field applied along direction of orientation; (*b*) field perpendicular to (*a*).

increases on increasing the applied field. Similarly, for a perpendicular field, the same type of highly nonlinear remanent-magnetization characteristic is obtained (Fig. 12-31*b*), but here lower maximum magnetization is achievable. Thus, taking account of the tape magnetization characteristics and the recording-head field contour, it is expected that the recording will be primarily longitudinal for most of the tape. However, for the surface layer in contact with the recording head, large perpendicular fields are encountered which produce some perpendicular magnetization. That this is the case has been verified by examining the tape magnetization direction at various depths into the magnetic layer with the use of a large scale model. It is an unfortunate feature of conventional recording that perpendicular recording, with relatively poor resolution, occurs in the surface layers where the maximum resolution would be obtained for longitudinal recording.

Digital-data Recording. The simplest method of coding the recording-head field is to reverse its direction. In digital-data recording a recording field of amplitude sufficient to produce magnetic saturation through the complete tape layer thickness is reversed to record a "1" signal and remains constant to record a "0" signal. Binary information recorded in this manner is known as *non-return-to-zero* (*NRZ*) *recording*, the most common technique for digital-data recording. Reproduction of this recording is achieved by using a timing signal, obtained from a separate

clock track, corresponding to the time when a 1 or 0 is recorded. Self-clocking systems are also in use where the recording field is reversed at regular intervals, and the 1 and 0 signals are recorded in between these clock signals.

The recording of an instantaneous reversal of the applied field will be considered initially for an infinitely thin layer of tape at a typical distance $y = 0.5g$ from the surface of the recording head. Considering the corresponding longitudinal field contour in Fig. 12-30, which is replotted in Fig. 12-32, and the longitudinal magnetization loops in Fig. 12-31*a*, the resulting tape magnetization on reversing the recording field can be computed for different maximum applied field amplitudes. The resulting form of the magnetization transition is shown in Fig. 12-32*a* for $H_{max} = 2H_s$ (solid line) and $H_{max} = 4H_s$ (dotted line). The derivative of the magnetization change for $K = 2$ is shown, in Fig. 12-32*b*, to be asymmetrical, and this corresponds to the reproduced waveform by use of an idealized reproducing head (time axis increasing to the left). As can be seen, both the width and the location of the reproduced pulse will change as the applied field maximum amplitude is increased.

With practical recording conditions, a number of factors will contribute to a widening of the reproduced pulse and a consequent decrease in the overall resolution. From the above construction of the reproduced pulse for an infinitesimally thin layer of tape, it can be expected that because of the change in field magnitudes and contours with distance from the head, the contribution from the other tape layers will occur with

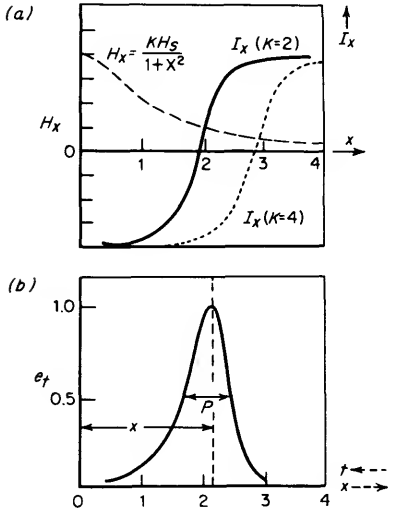


FIG 12-32 (*a*) Recorded transition; (*b*) ideal reproduction for NRZ step function.

different amplitudes and at different times. Also, the spatial sensitivity of the reproducing head is not an impulse function, as assumed above, but is given by the field function in Fig. 12-30. The reproducing-head flux is approximately expressed as a convolution integral of the longitudinal field function of the head H_x and the longitudinal tape magnetization I_x. For a tape of thickness c spaced a distance a from the head, the core flux ϕ_c is given by

$$\phi_c = K \int_a^{a+c} dy \int_{-\infty}^{+\infty} H_x I_x \, dx \tag{12-6-1}$$

where K is a constant. The time derivative of ϕ_c then gives the reproduced waveform. Further distortion of the pulse would be caused on taking account of the perpendicular component of magnetization, but this effect is relatively small. It is found, in practice, that even wider pulses than those estimated above are obtained, and self- and adjacent-bit-demagnetization effects must be considered to account for this. Such losses are reduced in very thin magnetic coatings with a relatively large ratio of H_c/I_x.

From the foregoing discussion of practical NRZ recording, it is evident that the highest resolution is obtained by adjusting the field amplitude so that the maximum longitudinal decrement occurs in the surface layer of the tape at a distance where the field has reduced to the switching fields of the particles in the tape material. In practice, larger fields are usually employed to ensure more reliable recording through the coating thickness. When faults occur in the tape coating, in the form of coating modules or foreign particles, the tape is momentarily separated from the recording head and a "dropout" of information occurs. To minimize the effect of dropouts, large recording fields are used and resolution is sacrificed for increased reliability. Present high-density data recording on oxide-powder tapes is in the range of 1,500 to 2,000 flux reversals per inch. By using thin metallic coatings with high coercive force, extension up to 10,000 reversals per inch is envisaged for the future. Corresponding improvements in the resolution of the reproducing heads is required, and this may be achieved by the use of narrow head gaps and electronic-pulse slimming filters.

Analog Recording. It is possible to extend the NRZ recording system to analog recording where amplitude linearity between the original and reproduced signals is required. This has been achieved by pulse coding the recording signal so that the width of the recorded pulses is controlled by the recording signal. Another system for analog recording with a tape-saturating signal is to frequency-modulate a high-frequency carrier, which is then recorded. However, by far the most economical system for analog recording uses direct recording of the signal, in which the amplitude and frequency of the signal are linearly recorded as changes of

magnetization amplitude and wavelength on the tape. A high-frequency ac-bias signal is added to the recording signal to achieve amplitude linearity. The resulting tape-magnetization process is rather complicated and has been described by numerous models. Fundamentally, the recording process with ac bias is a modification of ideal or anhysteretic magnetization, in which an alternating field, sufficient to saturate the material, is applied together with a constant field. On slowly reducing the alternating field to zero the magnetization achieved is approximately proportional to the dc field for magnetization levels up to about half the saturation level.

Anhysteretic Magnetization Process. The linearization of the nonlinear remanent magnetization curve I_r versus H, achieved in the anhysteretic magnetization process, is shown in Fig. 12-33 for oriented iron oxide–powder tape. For ac fields greater than 500 Oe a limiting condition is reached and the characteristic curve is obtained. In ac-bias recording, the anhysteretic magnetization process is modified, since the dc field (the recording signal) acting on a tape element reduces in amplitude at the same rate as the ac-bias field, as the element traverses the recording-head gap zone. This has little effect on the anhysteretic magnetization curves of Fig. 12-33 until the maximum ac field exceeds the coercive force. Then, instead of repeating the characteristic curve for large ac fields, the slope of the curve is reduced but keeps the same general shape. This is easily understood since the final magnetization of the particles is determined by the amplitude and direction of the dc field when the total field has reduced to the particle switching field H_c'. In anhysteretic magnetization this condition is always the same, providing the maximum ac field exceeds the particle switching field. However, in the modified magnetization process, the dc field occurring when $H_{ac} + H_{dc} = H_c'$ will be reduced as the maximum ac field increases. Thus, for recording conditions the initial anhysteretic susceptibility (or the recording sensitivity) has a maximum value for a maximum ac field approximately equal to the particle switching field. This corresponds to an optimum ac-bias amplitude.

It is observed that the general shape of the anhysteretic magnetization

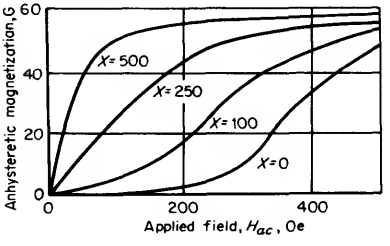


FIG 12-33 Anhysteretic remanent magnetization curves for oriented γ-Fe_2O_3 tape; x = peak ac field.

curve is similar for most hard magnetic materials. This shape reflects the distribution of internal fields acting on the particles. Whether or not a particle gets magnetized in the direction of an applied dc field during anhysteretic magnetization depends on whether H_{dc} exceeds the local internal field because of the magnetization of the neighboring particles. Essentially these local internal fields will be randomly directed throughout the material and will have an amplitude distribution which determines the shape of the anhysteretic magnetization curve. Locally, the magnitude and direction of the internal field will depend on the packing density of the particles and their magnetization directions. Consequently, if the magnetization direction of one particle is changed by an external field, the local field in its vicinity will also change and the effective switching fields of neighboring particles will be modified. One can see that the anhysteretic magnetization process does not then depend on the intrinsic particle switching fields; it depends on the internal fields and their effects on the particle switching fields. Owing to the angular dependence of these effects, somewhat higher susceptibility and linearity of the anhysteretic curve are obtained when the magnetizing fields are applied in the direction of orientation of the elongated particles, and this is the condition obtained in analog recording.

Practical Conditions. The recording process takes place in applied fields from a recording head having a distribution as shown in Fig. 12-30. This is more complicated than the modified anhysteretic magnetization process described above, since the magnetic coating, at different distances from the head, is subjected to fields of different amplitudes and directions. In addition, the recording field is somewhat modified by the fields from adjacent tape elements which have just been recorded. The net effect is that the actual recording process leads to a more linear characteristic than the modified anhysteretic magnetization curve. As the magnetic coating becomes relatively thicker, the long-wavelength recording sensitivity increases up to a point if the ac-bias amplitude is increased to give maximum sensitivity. Eventually, however, a limit is reached since a decrease of the magnetization of the surface layers in contact with the head occurs because of excessive ac-bias amplitude. In fact, examination of the recorded magnetization in large scale models indicates that very low magnetization levels are achieved in the surface layers. This is due in part to the low sensitivity to perpendicular fields of the longitudinally oriented tape and also to reduction of the recording field by fields from adjacent recorded zones.

In attempting to determine the optimum dimensions of the head gap and tape thickness, for broadband linear recording with ac bias, it is instructive to redraw the head field distribution in Fig. 12-30 as contours of constant field amplitude, as shown in Fig. 12-34. As has been

explained, ac-bias recording takes place when the total applied field $H_{ac} + H_{dc}$ falls to a value equal to the particle switching fields. Due to variations in these particle switching fields, amounting to about $0.25H_c$ for typical oxide powders, recording will take place over the finite region in which the applied fields fall into this range. Recording regions corresponding to three different amplitudes of applied field are shown as shaded zones in Fig. 12-34*a* and *b*. In Fig. 12-34*a* the zones correspond to the resultant applied field, and it is seen that the narrowest longitudinal recording region, giving highest short-wave length resolution, occurs for zone 1 where the field is not sufficient to magnetize the whole coating. This leads to nonoptimum recording to the long wavelength since the whole coating thickness can contribute to the long-wavelength output. Owing to separation losses in reproduction, only the surface layers contribute to the short-wavelength output, and remote layers are useless even if they could be recorded at short wavelengths. As can be seen, however, from Fig. 12-34*a*, if the bias amplitude is increased so that the whole of the coating is magnetized (zone 2), then the recording zone widens and the resolution of the recording deteriorates. If it were possible to produce a head in which the perpendicular field was attenuated, or a tape which was insensitive to this component, then the recording zone at the tape surface would not extend. This is illustrated in Fig. 12-34*b* where only the longitudinal component of the constant field contours is plotted. In practice, the short-wavelength recording resolution reduces so sharply with separation from the head gap that for a recorded wavelength of 2 μm, the loss caused by a 0.1-μm surface roughness on the tape is about 50 percent.

For the example shown in Fig. 12-34*a* and *b*, the recording-head gap

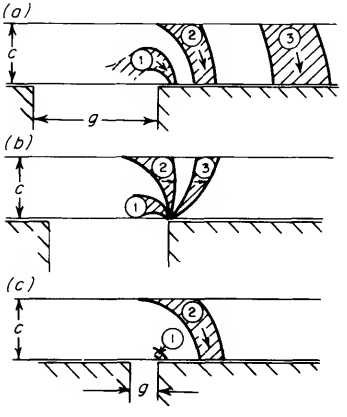


FIG 12-34 Constant recording-head field contours: (*a*) wide gap, resultant field; (*b*) wide gap, longitudinal field; (*c*) narrow gap, resultant field.

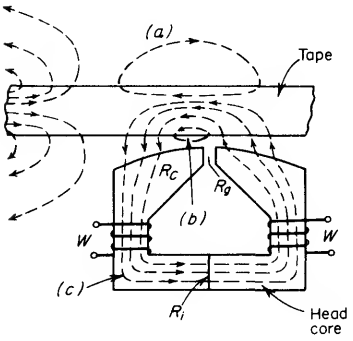


FIG 12-35 Schematic of reproduction process from a recorded tape.

length is twice the coating thickness. By using a narrow gap, as shown in Fig. 12-34*c*, correspondingly narrower recording zones are possible. However, because of the increased field decrement in the perpendicular direction, indicated by the small penetration of zone 1, the overall resolution is not improved if the field is increased to magnetize the whole tape as shown by zone 2. In practice, moderately narrow gaps are used to favor the recording resolution of short wavelengths. Thin coatings are also sought for the same reason. As the resolution requirements increase, the mechanical perfection of the gaps and of the head and tape surfaces becomes the limiting factor. Other losses can occur in the recording head itself, and thin metallic laminations or dense ferrites are used for the head core material to minimize the eddy current losses at the bias frequency. This frequency must be high enough to avoid any intermodulation with the signal and also to obtain anhysteretic magnetization conditions.

The Reproducing Process. A number of techniques are presently available for reproducing a magnetic recording. However, best transducing performance at moderate cost is obtained in an electromagnetic head with a high-permeability core. The essentials of the tape-reproducing function are shown schematically in Fig. 12-35. The low-reluctance core, having highly polished magnetic pole pieces in contact with the tape, shunts the flux external to the recorded tape through the desirable path C linking the head windings W. A small fraction of the flux closes through the useless paths a on the remote side of the tape and b between the head and the tape. If the recorded wavelength exceeds the length of the pole pieces in contact with the tape, a further loss occurs owing to flux not linking the core, and in the case of infinite recorded wavelength, the reproduced output falls to zero. The head will of course also deliver zero output if the frequency of a recording is reduced to zero by reducing the tape speed to zero. Under the normal reproducing con-

ditions shown in Fig. 12-35, most of the magnetic flux entering the head core from the tape takes the desirable low-reluctance path C consisting of the core reluctance R_c and the back gap reluctance R_i. However, the head gap provides a shunting reluctance $R_{g'}$, which may be low when the gap length is made very short to reproduce short recorded wavelengths. The fraction of the head core flux taking the desirable path C is given by

$$\frac{\Phi_c}{\Phi} = \frac{R_{g'}}{R_c + R_{g'} + R_i} \tag{12-6-2}$$

For high efficiency in the reproducing head, the gap reluctance is usually kept high by use of a very small gap depth, the disadvantage being the reduction in head life due to wear of the pole pieces. The core reluctance should be as small as possible, and high-permeability nickel-iron laminations are used for heads operating at relatively low frequencies. For video and pulse recording, however, the frequency losses in metal laminations are excessive, and ferrite heads are superior. Other practical reproduction losses associated with the head geometry arise when the pole-piece length is of the same order as the recorded wavelength and when the head gap either is not straight or is misaligned with the recording.

Reproduction of Analog Recording. The general technique used to calculate the reproducing-head flux due to a sinusoidal recording is to assume that since the reproducing process is essentially a linear one, the contribution of any tape element is proportional to the product of the magnetization of that element and the field which the head would produce at the element. In other words, a reciprocal relationship occurs, and the field distribution in the gap region may be looked upon as a measure of the spatial sensitivity of the head to the magnetization in the gap region. Integrating the contribution of all elements through the tape thickness and along the tape length yields the total head core flux Φ_c as shown in Eq. (12-6-1). For a finite gap length g', a tape thickness c, and a sinusoidal recorded magnetization I_x of wavelength,

$$\Phi_c = 4\pi c I_{x(\max)} ABC \cos \frac{2\pi vt}{\lambda} \tag{12-6-3}$$

where v is the tape velocity and A, B, and C are reproduction loss factors. Were it not for these losses, the reproducing-head voltage would be inversely proportional to the recorded wavelength; however, all three factors reduce the short-wavelength response.

Factor $A \left\{ = [1 - \exp(-2\pi c/\lambda)]/(2\pi c/\lambda) \right\}$ is called the *thickness loss* and refers to the attenuation of the reproducing-head flux compared with that obtained when the recorded wavelength is long with respect to the coating thickness.

Factor B $[= \exp(-2\pi a/\lambda)]$ also produces a monotonic decrease of the reproducing-head flux as the recorded wavelength is decreased. Here a is the head-to-tape spacing.

Factor C $(= [\sin(\pi g'/\lambda)/(\pi g'/\lambda)]\{[5 - 4(\lambda/g')^2]/[4 - 4(\lambda/g')^2]\})$ is a more complicated expression which accounts for interference effects in the gap region. As the recorded wavelength becomes shorter and approaches the gap length, destructive interference between oppositely magnetized elements of tape occurs with regard to their contributions to the head core flux. This leads to a series of minima in the core flux on further reducing the wavelength. Thus, in order to obtain a continuous wavelength response, it is necessary to use a sufficiently narrow gap so that the first minimum occurs outside the wavelength range of interest.

It is found that the calculated reproduction function of Eq. (12-6-3) agrees well with experimental results. It is also apparent that the losses occurring in recording with separation from the recording head are less severe than the reproduction losses described above. Some idea of the reproducing-head flux reduction due to separation loss for recorded elements inside the tape coating is obtained from the example that 75 percent of the output at the present limit of broadband recording is due to a coating thickness of only 0.37 μm. The importance of a smooth surface for the tape and heads is clearly shown.

Reproduction of Pulse Recording. In the process of reproducing the sharp tape-magnetization reversals that occur in pulse recordings, the objective is to obtain an output voltage spike having the smallest possible time spread. However, in practice the output pulse is widened by spreading of the magnetization transition of the recording and by the finite extent of the reproducing-head sensitivity function. In addition, interference from adjacent pulses causes the voltage spike to be shifted in time, which leads to possible errors in detecting the presence or absence of output signals. Normally, the outputs of a multitrack recording are sensed at a time specified by a separate timing track, and output pulses between 100 and 50 percent of maximum are counted as evidence of a recorded 1. In contrast to analog recording, it is not necessary to maintain linearity in the reproducing process, and the reproducing-head voltage pulses are often narrowed electronically to minimize interference effects; this is achieved by the use of filters which compensate for the low-pass filtering effects of the reproducing process.

Applying the reciprocity equation (12-6-1) to a recorded step function which is reproduced with a narrow-gap head, the output voltage for

infinitesimal coating thickness is given by

$$e_x = \frac{-n\,d\Phi_c}{dt} = KI_{x(\max)} \ln \frac{x^2 + (a + c)^2}{x^2 + a^2} \tag{12-6-4}$$

where K is a constant. This is in approximate agreement with the measured voltage pulses with the use of narrow-gap heads. Taking account of a small perpendicular magnetization component and of distortion in the recording process leads to the familiar asymmetrical shape of the reproduced pulse shown in Fig. 12-32*b*. Furthermore, taking account of a finite coating thickness and an effective head-to-tape separation leads to a widening of the reproduced pulse. At present, the use of very narrow reproducing-head gap lengths produces little resolution improvement since the recorded transitions are spread out by demagnetization effects in the tape. Again, tapes with very thin coatings and high ratios of H_c/I_r will approach the ideal step-function recording.

Noise in Reproduction. Any reproduced signal which is not part of the original recorded signal is defined here as noise. First, there are a number of unwanted signals arising from imperfections in the tape transport and in the recording and reproducing transducers; if the tape speed is caused to vary, either due to periodic changes in the transport system or due to oscillations set up by the friction between the moving tape and the stationary heads, undesirable changes in signal occur, which are known as "wow" and "flutter" respectively. Unwanted noise signals may of course be generated by the heads themselves or by the reproducing amplifier. However, the most noticeable noise signals usually originate in the tape itself.

When a demagnetized tape is reproduced, background noise may be detected. This could be due to insufficient demagnetization of the tape by the erasing head or, more fundamentally, to the particulate nature of the recording medium. The latter noise source is due to the randomly oriented, magnetically saturated domains which constitute the demagnetized tape. The noise voltage thus produced appears as if it were due to a noise magnetization of constant intensity over the whole frequency spectrum and is termed *white noise*. When the tape is magnetized, or becomes recorded with a signal, the noise behind the signal increases significantly above the background noise level. This is called *modulation noise* and may be 15 dB above the background noise level. It is caused by imperfections in the coating, such as surface asperities, as well as voids and agglomerations in the coating. Finally, unwanted signals can also be obtained from an effect known as *print-through*, which

occurs in a wound recorded reel of tape. During storage of a recorded tape, the magnetic fields of the recording penetrate the adjacent layers of tape in the reel and can sometimes produce a low-level print or recording which is detectable on reproduction. It is found that this undesirable effect increases with temperature and with the presence of small external fields. The single-domain particles of the recording medium have a distribution of shapes and sizes, and it is known that as the volume of such particles decreases, the relaxation times for magnetization change are reduced. In this case, the probability of magnetization in a small external field is increased, and the printed magnetization level is determined by the field, the ambient temperature, and the time of exposure.

Magnetic Recording Tapes. Although early developments in magnetic recording were made with flexible metallic tapes and wires as the recording media, the most successful type of recording medium for the last 20 years has consisted of a single thin magnetic coating on a flexible plastic tape. Typically, the plastic tape base is polyvinyl chloride (PVC) or polyethylene terephthalate, for example, *mylar*. Mylar tape is the preferred base material: it has inherent flexibility and immunity to humidity variations and is highly resistant to stretching and tearing. Furthermore, it can be manufactured with a minimum of surface defects which could cause corresponding defects (dropouts) in the magnetic coating. Almost universally the magnetic coating consists of a dispersion of very small particles of iron oxide γFe_2O_3 in a plastic binder. Typical tape dimensions are a width of 0.25 or 0.5 in. and thickness of 0.001 to 0.0005 in. The iron oxide particles are normally needle shaped, about 0.6 μm long and 0.1 μm in diameter; they occupy about 50 percent of the coating volume and are oriented with their long axes along the length of the tape. With a continuing emphasis on achieving higher recording densities, the surface smoothness of the coating and its ability to conform closely to the head contours become limiting factors. There is also a resolution advantage to be gained by using even thinner magnetic layers, and to a large extent, the reduction of magnetic material is offset by a higher intrinsic remanent magnetization in the layer. Consequently, higher coercive forces are required to avoid short-wavelength losses caused by self-demagnetization effects. Because of these requirements there has been a return to considering the advantages of metallic magnetic media. However, the metallic tapes now under development consist of a thin metallic layer deposited onto the plastic base material; in this way the physical and magnetic properties of the resulting tape are both optimum.

Magnetic Materials for Tape. From the consideration of the recording, reproducing, and storage conditions, it is evident that a highly nonlinear remanent magnetization is required, giving a low sensitivity to small

external fields. On the other hand, once the applied field magnitude reaches the irreversible magnetization threshold, a high-slope remanent magnetization is required for high recording sensitivity. An array of identical single-domain particles could fulfill these requirements. For instance, the switching field of a needle-shaped single-domain particle can be controlled by its shape if the switching fields because of other characteristics, such as crystal and strain anisotropy, are small. In this case, the preferred magnetization direction is along the length of the particle. When a field is applied opposite to the magnetization direction, no magnetization change occurs while the energy due to the applied field is less than the energy required to rotate the magnetization in the hard direction. When these energies are equal, the magnetization rotates completely in the field direction; thus, needle-shaped particles can yield the desired nonlinear magnetization characteristics. Crystal and strain energies may also be controlled to yield similar characteristics, but owing to thermal fluctuations the critical fields for switching are normally less stable than the shape-anisotropy-controlled particles.

In selecting a suitable material for shape-anisotropy-controlled single-domain particles, attention is paid to the size of range over which single-domain behavior exists. Outside this range, multidomain and superparamagnetic properties are obtained which reduce the magnetization stability. When the chosen particles are mixed with the plastic binder and spread onto the plastic base material, the aim is to produce complete orientation of the particles so that a recording field can be applied along their long axes to obtain the same magnetization characteristic in all particles. It is also desirable to produce a uniform dispersion so that the internal fields between particles have a low dispersion and so that noise due to particle clumping is minimized. The extent to which practical particle dispersions fulfill the above requirement is shown in Fig. 12-36; the needle-shaped iron oxide particles, γFe_2O_3, are typical of present-day tape-recording media. The saturation magnetization I_s of such a tape is about 160 G, and $I_r/I_s = 0.75$ for practical oriented tapes. The coercive force $H_c = 250$ Oe has a magnitude corresponding to an incoherent magnetization rotation process. Hysteresis loops and anhysteretic-magnetization curves in the direction of particle orientation are shown for this material in Figs. 12-31 and 12-33 respectively. Other oxide powders tried in the past for magnetic tapes include small particles of cobalt-doped iron oxide in which crystal anisotropy dominates. Higher coercivities are thereby obtained which reduce the tendency to self-demagnetization at short wavelengths. However, the large temperature variation of crystal anisotropy produced poor storage stability in this material.

Needle-shaped metal powders, on the other hand, do appear at first

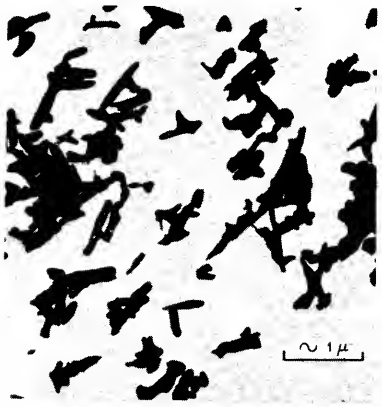


FIG 12-36 Horizontally oriented iron oxide particles, magnified 20,000 X.

sight to offer good tape material characteristics. The single-domain size is very small (0.04-μm length), promising low noise. Tape saturation magnetization and coercivity, on the other hand, are high compared with the oxides; for instance, for iron particles, $I_s = 680$ G and $H_c = 850$ Oe. Thus, high-output tapes are possible or, alternatively, very thin layers may be used. Unfortunately, the latter advantage cannot be fully realized, since coating techniques cannot produce satisfactory layers less than 2 μm thick. Compared with oxides, self-demagnetization effects will not be reduced in iron particles having the above properties. However, higher coercivities have been obtained in cobalt particles where coherent magnetization rotation processes have been observed. Alloy particles of iron-cobalt and cobalt-nickel also show somewhat similar properties indicating at least a threefold increase in remanent magnetization compared with oxide powders.

Another approach to the manufacture of very thin metallic magnetic-layer tapes has been attempted by using electrodeposition or electrodeless deposition of the material onto the plastic base material. In this method the inherent advantage of coating dispersions is lost; that is to say, there is less room for manufacturing error, since no averaging out of variations in magnetic properties can take place. Such layers of cobalt-nickel-phosphorus (Co-Ni-P) have been used for a number of years on magnetic drums and are now successfully deposited onto a mylar base. By using very thin layers of Co-Ni-P or Co-P (say, 0.2 μm), the flexible properties of the mylar are not impaired and recording advantages are obtained. All the coating thickness contributes to the reproducing-head signal even for the shortest wavelengths. Also, self-demagnetization and adjacent-bit-demagnetization effects are reduced for such thin layers. The loss in output due to the use of thin metal films is largely offset by the high

intensity of magnetization, I_s = 900 G, and by the reduced losses described. It thus appears that thin-metallic-film tapes have the potential to replace oxide tapes providing an adequate degree of quality control can be maintained in their production.

CITED REFERENCES

1. Davis, Sidney A., and Byron K. Ledgerwood: "Electromechanical Components for Servomechanisms," McGraw-Hill Book Company, New York, 1961.
2. Davies, G. L.: "Magnetic Tape Instrumentation," McGraw-Hill Book Company, New York, 1961.
3. Hoagland, A. S.: "Digital Magnetic Recording," John Wiley & Sons, Inc., New York, 1964.
4. Mee, C. D.: "The Physics of Magnetic Recording," North-Holland Publishing Company, Amsterdam, 1964.
5. Spratt, H. C. M.: "Magnetic Tape Recording," Macmillan Company, New York, 1964.
6. Winckel, F.: "Technik der Magnetspeicher," Springer-Verlag OHG, Berlin, 1960.

CHAPTER THIRTEEN

MEASUREMENTS ON AUDIO AND VIDEO AMPLIFIERS

From Notes by

Paul Baird

Fred Hanson

Charles Kingsford-Smith

Terry E. Tuttle and

Larry A. Whatley

Hewlett-Packard Company, Loveland, Colorado

All the instruments and measurement techniques described in other chapters are used from time to time on amplifiers, but it is desirable to study amplifier characteristics further. Enough theory and design considerations will be given in this chapter to clarify the various measurements treated. For a fuller treatment of amplifier calculations the reader is directed to a good text [1, 2]. Familiarity with Chap. 2 is assumed.

As stated in Chap. 7, amplifiers serve (1) to increase the power avail-

able in an electrical signal, (2) to amplify voltage or current levels where power per se is not of great concern, (3) to *limit* automatically the voltage or current that can be delivered to a load, (4) to provide a prescribed transfer function, either linear or nonlinear, between a source and a load, (5) to provide the desired load impedance on a source or the desired source impedance for a load, and (6) to attenuate or reject the common-mode component of voltage on a pair of conductors (the common-mode voltage is the *average* of the voltages on the two conductors at each instant with respect to ground or some designated reference potential).

13-1 Transfer Gain and Transfer Function

In this chapter, the transfer gain of an amplifier is defined as the ratio of an output quantity to an input quantity. Gain can be expressed as a ratio of amplitudes or as a ratio of amplitudes with associated phase differences, and both amplitude and phase difference are generally functions of frequency. When the ratio of output to input is expressed as a function of the complex frequency s, the ratio is known as the *transfer function.*

Both the input and the output variables in an amplifier can be considered either voltage or current. This leads to four possible gain ratios: voltage gain, current gain, transconductance, and transimpedance. Refer to Fig. 13-1 for the basic circuits of an amplifier with Thevenin and Norton equivalents for the signal source. Regardless of how gain is defined for these circuits, it is usually desired that the gain be insensitive to changes in Z_s, Z_L, and environmental conditions. For instance, if the amplifier is used for voltage gain E_2/E_s, it is desirable that $Z_i \gg Z_s$ and $Z_o \ll Z_L$ to keep changes in these impedances

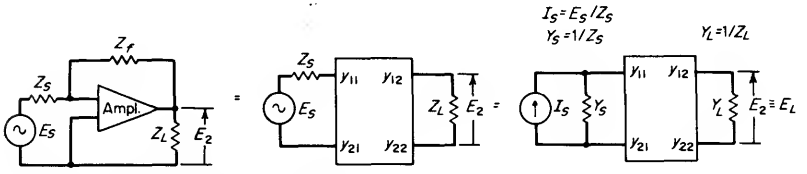


FIG 13-1 Operational shunt feedback.

from affecting the gain ratio. The desired impedance relationships are achieved by amplifier design, often through the use of feedback.

Other desired impedance relationships are shown in Table 13-1, along with the suggested network parameter for analysis [1].

When the chosen transfer gain needs to be insensitive to internal param-

TABLE 13-1 Impedance Measurements with Amplifiers

Transfer gain	Preferred Z_{in}	Preferred Z_o	Convenient parameters for analysis
Voltage gain E_L/E_s	$\gg\|Z_s\|$	$\ll\|Z_L\|$	h
Current gain I_L/I_s	$\ll\|Z_s\|$	$\gg\|Z_L\|$	g
Transconductance I_L/E_s	$\gg\|Z_s\|$	$\gg\|Z_L\|$	z
Transimpedance E_L/I_s	$\ll\|Z_s\|$	$\ll\|Z_L\|$	y

eter variations as well as source and load terminations, a suitable type of feedback is employed. For each of the four kinds of gain in Table 13-1 there is a feedback arrangement which will shift input and output impedances in the desired directions and desensitize the amplifier to internal parameter variations.

Consider, for example, the operational (or shunt) type of feedback, Fig. 13-1*a*. The amplifier voltage gain is inverting and over a range of frequencies $E_L/E_s \approx -Z_f/Z_s$. Here we have just expressed a voltage gain, which is commonly done.

Considerable insight can be gained, however, by viewing the amplifier as providing a transimpedance rather than a voltage gain. The voltage-inverting amplifier, Fig. 13-1*a*, combined with the shunt impedance Z_f can be viewed as forming a new amplifier which has very low input and output impedances. The y parameters of this new amplifier can be determined by adding term by term the parameters of the original amplifier to the y parameters of the shunt impedance. For this reason and for simplicity in discussing the ratio E_L/I_s, the y parameters are convenient for this particular viewpoint. The transfer function E_L/I_s is insensitive to load and source variations, while E_L/E_s is quite sensitive to source variations.

One of the reasons for making loop gain measurements on amplifiers (in addition to shaping frequency response for stability) is that loop gain is closely related to sensitivities. If a transfer gain is denoted by a symbol T, the sensitivity of T with respect to a parameter k (denoted S_k^T) is defined by

$$S_k^T = \frac{dT/T}{dk/k} \tag{13-1-1}$$

The sensitivity of T to k is the fractional change in T divided by the fractional change in k (which caused the change in T), with all fractional changes considered infinitesimally small.

For the case of

$$\frac{E_L}{I_s} = \frac{-y_{21}}{(y_{11} + Y_s)(y_{22} + Y_L) - y_{12}y_{21}} \tag{13-1-2}$$

$$S_{y_{21}}^{E_L/I_s} = \frac{1}{1 - y_{12}y_{21}/[(y_{11} + Y_s)(y_{22} + Y_L)]} \tag{13-1-3}$$

The loop gain of concern here (it is possible to define various loop gains) is $y_{12}y_{21}/[(y_{11} + Y_s)(y_{22} + Y_L)]$. For those who prefer to discuss amplifiers in terms of K and β, we can set $E_L/I_s = K/(1 - \beta K)$ and

$$S_{y_{21}}^{E_L/I_s} = \frac{1}{1 - \beta K}$$

by identifying

$$K \equiv \frac{-y_{21}}{(y_{11} + Y_s)(y_{22} + Y_L)} \qquad \text{ohms}$$

and $\beta \equiv -y_{12}$ mho. As a practical matter most laboratory equipment measures voltages, and so βK would be measured as a voltage ratio. Figure 13-2 shows how a floating voltage source, such as the secondary of a shielded transformer, can be inserted into the feedback loop to enable one to measure loop gain. Here, if $|Y_f| \ll |Y_L +$ amplifier $Y_{22}|$, then

$$\frac{E_L}{E_A} = K\beta = \frac{y_{12}y_{21}}{(y_{11} + Y_s)(y_{22} + Y_L)} \tag{13-1-4}$$

Regardless of what feedback arrangement is used to produce gain stability and the various impedance relationships shown in Table 13-1,

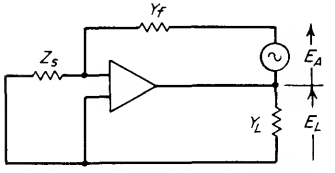


FIG 13-2 Circuit for measurement of loop gain in an amplifier with shunt feedback.

it is usually convenient to express loop gain in a manner that allows it to be found with voltage measurements rather than current measurements, although current can sometimes be measured in conductors by means of clip-on current instruments.

13-2 Steady-state Gain and Phase Measurement in Amplifiers

While the emphasis in this section is on amplifier measurement, the techniques are also applicable to passive transmission networks. Linearity is assumed in the amplifiers or networks. In other words, the output signal $E_o(j\omega)$, the input signal $E_i(j\omega)$, and the ratio of the two signals have specific values of amplitude and phase angle at any given ω. In consistency with the whole chapter, it will be understood that signals are functions of $j\omega$, and the $(j\omega)$ will be dropped from the text symbols.

Figure 13-3 shows the most direct way to measure the amplitude and phase of the gain ratio. The signal source is adjusted to the desired frequency and amplitude, and then the phase meter and the two ac voltmeters are read. The voltage gain is calculated as

$$K = \frac{|E_o|}{|E_i|} \underline{/\phi_o - \phi_i} \tag{13-2-1}$$

where $\phi_o - \phi_i$ is the difference between output and input phase angles.

The phase meter can be either a digital or an analog instrument. Chapter 6 discusses digital phase meters, and analog phase meters measure the ratio of the interval between zero crossings of two signals to the period of one cycle of the test frequency. See p. 25 for general treatment of phase detectors.

When the components of the test arrangement in Fig. 13-3 are combined into a single instrument, it is called a *network analyzer*. The one shown in Fig. 13-3 would be considered an elementary analyzer; it is limited in accuracy by the frequency responses of both the voltmeters and the phase meter.

A more accurate and versatile network analyzer is shown in Fig. 13-4. It has two prominent features. First, the test signal is heterodyned to a constant frequency $\Delta\omega$ at the mixer output before it is measured, regardless of the frequency at which the amplifier or network is being tested. Second, the input signal is heterodyned and filtered identically before it is used as a reference.

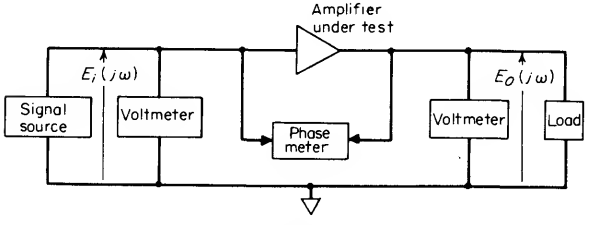


FIG 13-3 Direct method for measurement of complex gain.

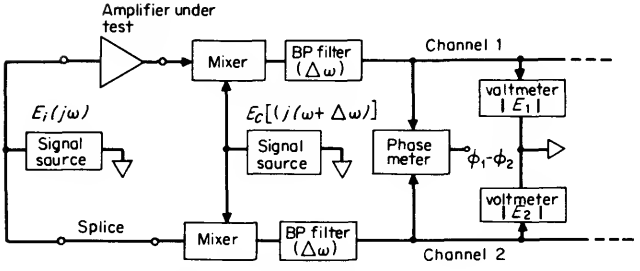


FIG 13-4 Dual-channel network analyzer.

In Fig. 13-4, the network to be tested is inserted in channel 1, and channel 2 is bridged with a short cable. Both channels are fed into high-impedance frequency converters or mixers. A third generator, offset by a constant frequency from the other two, provides the switching signal for both mixers. The identical bandpass filters at the outputs of the mixers select only the difference frequency signals, in which are preserved both the phase and the amplitude of the mixer input signals. Now, if the channels are identical, the input voltage to mixer 2 is a duplicate of that applied to the network in channel 1. Hence the complex ratio E_1/E_2 is the desired voltage-gain transfer function.

It is usually not necessary to measure the voltage ratio $|E_1|/|E_2|$, since if E_1 is constant, $|E_2|$ is also constant and thus the magnitude of K is proportional to $|E_1|$.

For good accuracy, it is desirable that the two channels be as nearly identical as possible. If small differences exist that are constant, they can be kept from affecting the accuracy by means of a calibration technique. In other words, the output voltages and phase difference are noted (and perhaps adjusted) with the network under test replaced by a short length of transmission line.

Amplifiers with Feedback. The above remarks apply, of course, to amplifier circuits, which are treated as linear networks. However, amplifier circuits often include the special consideration of feedback, which makes the network analyzer a very valuable tool for design analysis and modification. The relation between open- and closed-loop characteristics is well understood in amplifiers partly because of work in the field of feedback control systems. The designer of feedback amplifiers often makes more use of experimentally determined characteristics than his counterpart working with control systems. There are several related reasons for this:

1. Operating frequencies in amplifiers are often much higher, and circuit parameters are more difficult to determine analytically to the required degree of precision.

2. Stray and parasitic elements become important at these higher frequencies.

3. Signal transit time through an amplifier results in considerable phase shift at high frequencies.

When operating with large signals, both feedback amplifiers and control systems sometimes exhibit a variation in gain over the signal dynamic range, a variation caused, for instance, by the change of grid-plate transconductance as the plate current varies in a vacuum tube. This may result in instability over a portion of the signal waveform. A network analyzer, used to define the incremental gain and phase as the system is exercised over its dynamic range, is a very useful aid in determining the required compensation.

It is worthwhile noting that reason 3 above is often responsible for unexpected closed-loop characteristics. The designer expects that the gain-versus-frequency function of the open loop—obtained with the use of a swept-frequency generator, or point by point—will accurately predict the closed-loop performance. This is usually the case in simple control systems, for they are usually minimum-phase networks with phase shift established uniquely by the gain characteristic.

However, the excess phase resulting from transit time or nonminimum phase conditions may render the closed-loop performance unstable or unacceptably peaked. A network analyzer, which measures *both* gain and phase components of the open-loop characteristic, enables the designer to determine accurately the closed-loop performance.

Bode Plots. Network analyzers are available that provide a swept-frequency source and outputs proportional to the phase and log amplitude of the network transfer function. Such network analyzers are easily used for making Bode plots [3] automatically. An oscilloscope or *xy* recorder is used as the display device.

To measure the phase and gain margins, the network analyzer is first calibrated with a cable in place of the amplifier network. This establishes reference levels for both phase and gain. The amplifier is then connected and its characteristic curve plotted. The required stability margins can be read directly from the plot.

Figure 13-5 shows the Bode plot of the open-loop response of an experimental 10-MHz video amplifier with total feedback. In this case, the engineer had a particular amplifier on hand and wished to achieve high input impedance, low output impedance, high power gain, and good gain stability by feeding all the output back in series with the input. The resulting voltage gain of approximately unity was acceptable. Originally there were two negative real poles at 1 MHz and a single negative pole at 100 MHz. The designer inserted a lead-lag correction network with a real axis zero and pole at -4 and -20 MHz, respectively. The gain

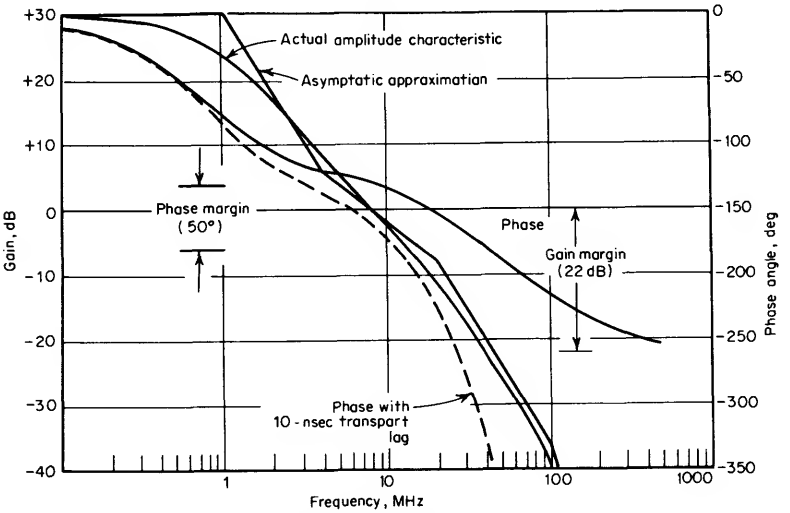


FIG 13-5 Bode plot of an experimental amplifier.

and phase margins achieved thereby are satisfactory. However, the existence of a 10-nsec transport lag, discovered by means of a network analyzer, reduces both margins to unacceptable levels. Naturally, such degradation of the closed-loop response is not evident from the amplitude plot alone.

Nichols Plots [4]. While a Bode plot of the open-loop characteristic gives much insight into closed-loop behavior, it has the disadvantage of not providing direct quantitative data about the closed-loop response. A Nichols plot overcomes this limitation. By using a network analyzer to generate data for a Nichols plot of an open-loop response, the designer can adjust circuit parameters and immediately observe the effect on the closed-loop response.

A Nichols chart (Fig. 13-6) is an xy graph with phase shift as abscissa and log of gain as ordinate. Superimposed on the graph are two sets of loci: One set represents various values of closed-loop gain; the other, values of closed-loop phase shift. When the open-loop characteristics of an amplifier are plotted on such a graph, with frequency as a parameter, intersections with the constant gain and phase loci indicate the closed-loop behavior.

A Nichols plot of the amplifier with the Bode plot of Fig. 13-5 is reproduced in Fig. 13-6. The solid trace represents the compensated loop, while the dotted trace includes the effect of a 10-nsec transport lag. Notice that the actual closed-loop performance can be read from the

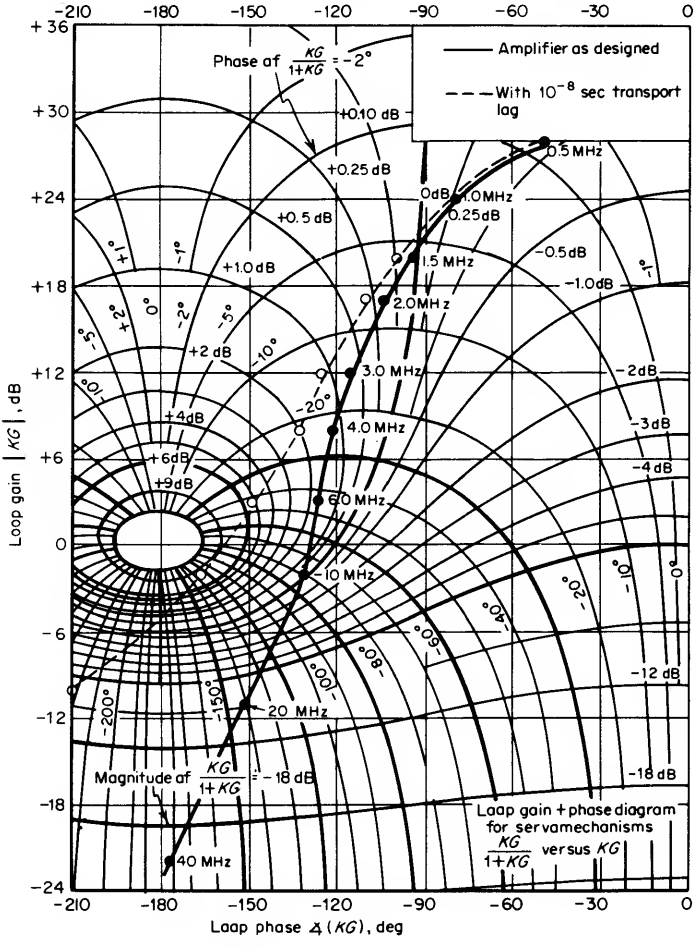


FIG 13-6 The Nichols chart.

chart: 1.6 dB peaking at 6 MHz. Likewise, with the transport lag included, the closed-loop response peaks about 10 dB at 9 MHz.

One-port Immittances. Although the network analyzer has been described as a transfer measuring device, it is easy to adapt it to measure input and output characteristics of an amplifier (or of any linear one-port device). A common technique is to convert the variable-frequency generator into a constant-current source (for impedance measurement) or a constant-voltage source (for admittance measurements)

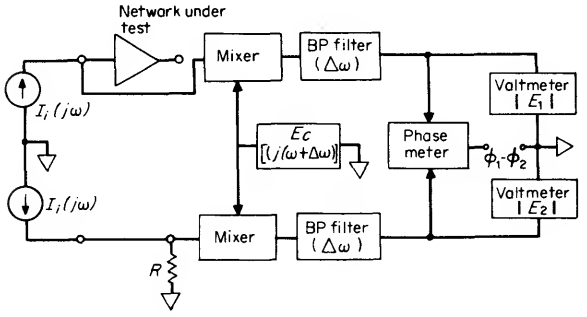


FIG 13-7 Two-terminal impedance measurement.

to drive the network under test. The analyzer then measures the complex voltage (or current) response, which is proportional to the impedance (or admittance) of the one-port network.

Figure 13-7 shows the implementation of this technique, with a dual-channel network analyzer to measure impedance. As before, it is not necessary to measure the voltage ratio $|E_1|/|E_2|$ if constant-current generators are used, since $|E_2|$ will be constant. Calibration of this *complex ohmmeter* is readily achieved by inserting a known resistance in place of the network under test in channel 1.

13-3 Gain Measurement with Extreme Accuracy

The instruments and procedures described in the preceding section are suitable when gain measurement with an error not lower than about 0.1 percent is acceptable. Some systems require amplifiers with lower gain errors than 0.1 percent, however, and special techniques must be used to measure gain with such accuracy. A case is described below where it was desired to measure the voltage gain of a noninverting amplifier with accuracies to 0.01 percent over a 100-kHz range.

Figure 13-8 is a diagram of the scheme that was used. The simple phase shifter is variable over only a few electrical degrees, and it can be

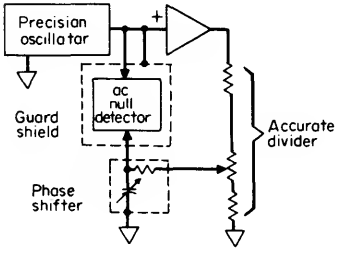


FIG 13-8 Arrangement to measure gain accurately.

switched to either side of the null detector, depending upon the sense of the amplifier phase error. To use this setup, the resistive divider and phase shifter are adjusted until a null is obtained to the desired sensitivity. The resistive divider is then removed, and its ratio is measured with dc. This ratio equals the gain (except for a phase-shifter correction). The frequency range is limited primarily by the frequency response of the resistive divider, since it is not pure resistance. The method works best when the phase shift required for a null is very small. This is usually the case if the amplifier gain is very accurate and the divider frequency response is good enough to check it. Table 13-2 tabulates amplitude corrections to be made for various amounts of phase shift. The amplitude at the output of the phase shifter (across the capacitor) is slightly less than at its input.

Note that neither side of the null detector can be grounded for this setup. To achieve freedom from power-line interference and to get the required CMR, a battery-operated null detector surrounded by a guard box is shown. If the system should be further modified by using an input divider between the oscillator and amplifier, a second input divider would

TABLE 13-2 Amplitude Correction Required by Phase-shifting Network Shown in Fig. 13-8†

Angle, deg.	*Correction, %*
10	1.5192247
9	1.2311659
8	0.973193
7	0.7453848
6	0.5478
5	0.3805
4	0.2436
3	0.1370
2	0.0609
1	0.0152
0.9	0.01234
0.8	0.00975
0.7	0.00746
0.6	0.00548
0.5	0.00381
0.4	0.00244
0.3	0.00137
0.2	0.00061
0.1	0.000152

† Phase-shifter output is less than its input by the amount stated.

then be added to drive the guard box separately. Since no input divider has been shown, the guard box is correctly driven.

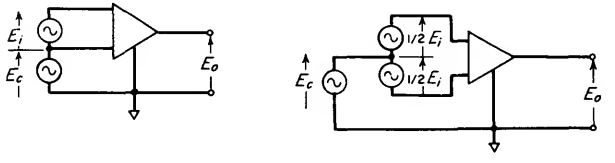
Even better accuracy than cited above can be obtained over a narrow frequency range by substituting a suitable inductive divider for the resistive one. Inductive dividers are available that give a direct readout of ratio, and accuracies to 0.0001 percent are possible.

In all cases above, it is assumed that the distortion of the amplifier is very low. Otherwise a sharp null could not be obtained, at least not with a broadband null detector.

13-4 Common-mode Rejection and Input Balance

The signal in a network that one wishes to measure is frequently combined in some way with an offending signal, and the problem is to measure the first in the presence of the other. Sometimes the offending signal is simply superimposed upon, or added to, the desired signal. When this series condition exists, separation can only be made by filtering, and filtering requires a priori knowledge of the frequency content and time relationship of the two signals. Of course, the interfering signal can often be reduced or eliminated at its source or by shielding.

The other kind of interfering signal is called *common mode* because it is common to both input terminals of an amplifier, as in Fig. 13-9. For



(a) Unbalanced (b) Balanced

FIG 13-9 Signal and common-mode inputs to a differential amplifier.

either the balanced or unbalanced equivalent circuit,

$$E_o = K_d E_i + K_c E_c \tag{13-4-1}$$

where K_d is the differential gain of the amplifier, defined as

$$K_d = \frac{E_o}{E_i} \qquad E_c = 0 \tag{13-4-2}$$

and K_c, the common-mode gain, is

$$K_c = \frac{E_o}{E_c} \qquad E_i = 0 \tag{13-4-3}$$

In other words, the differential amplifier is not perfect, and in general, a common-mode input signal produces a small but finite output superimposed upon the useful signal.

The CMR ratio of the amplifier is defined here as

$$\text{CMR}_A = \frac{K_c}{K_d} \tag{13-4-4}$$

Stated another way, it is the ratio of the input signal to the common-mode signal that produces equal components of output. This definition generally gives $\text{CMR}_A \ll 1$. Expressed in decibels, it gives a fairly large negative number. Some other writers have chosen to use K_d/K_c as CMR, or CMR ratio, but as long as we are consistent here, there should be no confusion. A good CMR ratio goes along with a small fraction, not a large number.

For small signals, one can use either *a* or *b* of Fig. 13-9 to represent the operating conditions. However, when source impedances and amplifier input impedances are considered, the equivalent circuit becomes more complex, as shown in Fig. 13-10. Here, E_c can in general produce differential components of signals at the input and output of the amplifier, even if CMR_A is zero. The CMR ratio that the whole circuit would have if CMR_A were zero is called *balance*. The symbol is CMR_B. In the figure, Z_{s1}, Z_{s2}, and Z_{s3} are impedances associated with the signal source, and Z_i, Z_1, and Z_2 are characteristic of the input to the amplifier.

Referring to Fig. 13-10, observe that E_o actually consists of four components, which arise from the following causes:

1. The differential gain N_{dd} of the input network and differential gain K_d of the amplifier,

$$E_{o1} = N_{dd}K_dE_i$$

2. The differential–to–common-mode gain N_{dc} of the input network

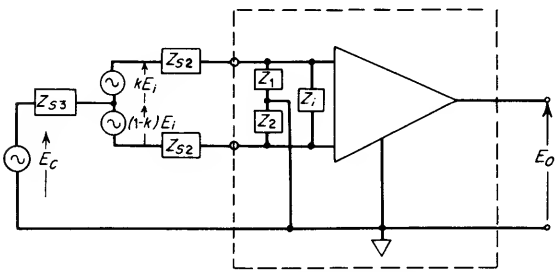


FIG 13-10 Circuit to illustrate balance and CMR.

and the common-mode gain K_c of the amplifier,

$$E_{o2} = N_{dc}K_cE_i$$

3. The common-mode–to–differential gain N_{cd} of the input circuit and the differential gain K_d of the amplifier,

$$E_{o3} = N_{cd}K_dE_c$$

4. The common-mode gain N_{cc} of the input circuit and the common-mode gain K_c of the amplifier,

$$E_{o4} = N_{cc}K_cE_c$$

Output components 1 and 2 are due to differential input signal, while 3 and 4 are due to common-mode input. Adding like terms,

$$E_{o12} = N_{dd}\left(K_d + \frac{N_{dc}}{N_{dd}}K_c\right)E_i \tag{13-4-5}$$

$$E_{o34} = N_{cd}\left(K_d + \frac{N_{cc}}{N_{cd}}K_c\right)E_c \tag{13-4-6}$$

Since E_{o12}/E_i is the differential gain of the system and E_{o34}/E_c is the common-mode gain, the CMR of the whole system is

$$\text{CMR} = \text{CMR}_A \cdot \text{CMR}_B\left[\frac{1/\text{CMR}_A + N_{cc}/N_{cd}}{1 + \text{CMR}_A(N_{dc}/N_{dd})}\right] \tag{13-4-7}$$

where CMR_A = CMR of $A = K_c/K_d$, and CMR_B = CMR of

$$N = \frac{N_{cd}}{N_{dd}}$$

Equation (13-4-7) shows that CMR is not just the product of CMR_A and CMR_B. Under the condition that A has ideal CMR, that is, $\text{CMR}_A \to 0$,

$$\text{CMR} = \text{balance} = \text{CMR}_B = \frac{N_{cd}}{N_{dd}} \tag{13-4-8}$$

Equation (13-4-7) may be written

$$\text{CMR} = \text{CMR}_B\left[\frac{1 + (N_{cc}/N_{cd})\text{CMR}_A}{1 + (N_{dc}/N_{dd})\text{CMR}_A}\right] \tag{13-4-9}$$

We recall that N_{cc}/N_{cd} is the ratio of common-mode–to–differential output for common-mode input of the input network. Also, N_{dc}/N_{dd} is the ratio of common-mode–to–differential output for a differential input. Redrawing the input network N from Fig. 13-10, we have Fig. 13-11.

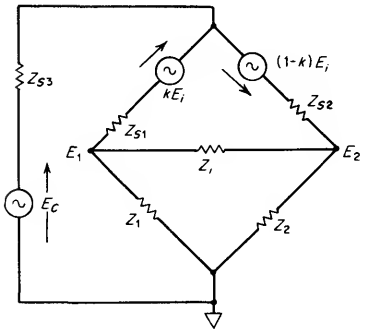


FIG 13-11 Input network of Fig. 13-10.

By definition,

$$\frac{N_{cc}}{N_{cd}} = \frac{[(E_1 + E_2)/2]/E_c}{(E_1 - E_2)/E_c} = \frac{E_1 + E_2}{2(E_1 - E_2)}\bigg|_{E_c=\text{input}}$$
$$= \frac{Z_{s1}Z_2 + Z_{s2}Z_1 + 2Z_1Z_2(Z_{s1} + Z_{s2} + Z_i)/Z_i}{2(Z_{s2}Z_1 - Z_{s1}Z_2)} \tag{13-4-10}$$

and

$$\frac{N_{dc}}{N_{dd}} = \frac{[(E_1 + E_2)/2]/E_i}{(E_1 - E_2)/E_i} = \frac{E_1 + E_2}{2(E_1 - E_2)}\bigg|_{E_i=\text{input}}$$
$$= \frac{(1 - 2k)Z_1Z_2 + Z_{s3}(Z_2 - Z_1) + (1 - k)Z_{s1}Z_2 - kZ_{s2}Z_1 + 2Z_1Z_2[(1 - k)Z_{s1} - kZ_{s2}]/Z_i}{-2[Z_{s1}Z_2(1 - k) + kZ_{s2}Z_1 + Z_{s3}(Z_1 + Z_2) + Z_1Z_2]} \tag{13-4-11}$$

The balance is

$$\frac{N_{cd}}{N_{dd}} = \frac{Z_{s1}Z_2 - Z_{s2}Z_1}{(1 - k)Z_{s1}Z_2 + kZ_{s2}Z_1 + Z_{s3}(Z_1 + Z_2) + Z_1Z_2} \tag{13-4-12}$$

In a balanced transmission line with matched impedances,

$$Z_{s1} = Z_{s2} = Z_{s3} = \frac{R_0}{2}$$

and $Z_i = R_0$, where R_0 is the characteristic impedance of the line. Also, $k = \frac{1}{2}$. Under these conditions, one can use the above equations to find that balance is

$$\text{CMR}_B = \frac{N_{cd}}{N_{dd}} \approx \frac{2(Z_2 - Z_1)}{3(Z_1 + Z_2) + 4(Z_1Z_2)/R_0} \tag{13-4-13}$$

If we further assume $Z_1 \gg R_0 \ll Z_2$,

$$\text{Balance} \approx \frac{(Z_2 - Z_1)R_0}{2Z_1Z_2} \tag{13-4-14}$$

In this situation, the CMR of the whole circuit is

$$\text{CMR} \approx \text{balance}\,\frac{1 + [4Z_1Z_2/R_0(Z_1 - Z_2)]\text{CMR}_A}{1 + [3Z(Z_1 - Z_2)R_0/8Z_1Z_2]\text{CMR}_A} \tag{13-4-15}$$

For $\text{CMR}_A \ll 1$,

$$\begin{aligned}\text{CMR} &\approx \text{CMR}_B\left(1 - \frac{2\ \text{CMR}_A}{\text{balance}}\right)\\ &= \text{CMR}_B - 2\ \text{CMR}_A\end{aligned} \tag{13-4-16}$$

or the balance is

$$\text{CMR}_B \approx \text{CMR} + 2\ \text{CMR}_A \tag{13-4-17}$$

Since the output voltage produced by a common-mode input can be either inverted or not, CMR and CMR_A can have either the same sign or different sign, and to emphasize this we write

$$\text{Balance} = \text{CMR} \pm \text{CMR}_A \tag{13-4-18}$$

under the requirements that $\text{CMR}_A \ll 1$ and R_1, R_2, $R_3 \ll Z_1$, Z_2.

The relationship between CMR and balance has been shown; Eq. (13-4-9) shows the exact expression relating balance and CMR, while Eq. (13-4-18) shows the simple approximate relationship for the transmission-line example.

Low CMR and balance may be attained by several methods. One of these is the use of a transformer. Another method uses the differential amplifier with active components.

13-5 Common-mode Rejection and Balance with Transformer Coupling

Figure 13-12 shows how a shielded transformer is connected between a signal source and an amplifier to reduce common-mode output. The impedances to ground, Z_1 and Z_2, comprising R_1 and C_1 and R_2 and C_2, consist of insulation resistance and winding capacitance, and the impedance of these elements can approach $10^9\ \Omega$ at low frequencies. The Z_i of the earlier derivations is Z_L reflected to the primary in accordance with the turns ratio of the transformer. Therefore, in the lower frequency range, obtaining good (-80 to -90 dB) CMR and balance is not difficult.

However, at higher frequencies, the stray capacitance of the system

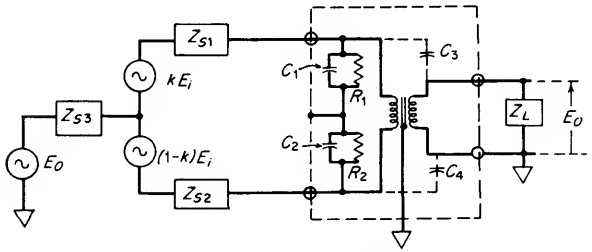


FIG 13-12 Transformer used to reduce CMR.

plays a very important role and must be given careful consideration. Let C_1 and C_2 represent primary-winding capacitance between the wire and the core. The capacitors C_3 and C_4 will effectively modify the values of C_1 and C_2 respectively. But C_3 does more damage than just adding to C_1 since it couples the common-mode signal E_c directly to the output.

This capacitance alone keeps the common-mode gain E_o/E_c from being reduced much below $Z_L/(Z_L + 1/j\omega C_3)$. However, if a good grounded electrostatic shield is interposed between the primary and secondary windings, C_3 can be reduced almost to zero. Also, transformers can be wound in sections and interconnected in a manner that makes the equivalent values of C_1 and C_2 almost equal, and this gives good balance over a wide range of frequencies.

The use of a transformer limits the frequency response, although operation over a frequency range of three decades with errors of less than ½ dB is readily attainable. One very important advantage of a transformer is the large common-mode voltage capability, often in the neighborhood of several hundred rms volts.

13-6 Differential Amplifiers with Low CMR and Balance Ratios

Obtaining good CMR and balance with differential amplifier techniques does not come as easy as with a transformer, but accuracy of gain and width of frequency are far superior. Generally, balance at higher frequencies is much better because of the lower and more tightly controlled input capacitance.

One of the better methods for ensuring good CMR is to use a compensating common-mode amplifier. This is a single-ended amplifier which amplifies only the common-mode signal present at the input. The output of the amplifier is used to drive one or more points in the main amplifier including, perhaps, an enclosing box. The objective is to allow amplification of differential signals with minimum common-mode–to–differential response. Rejection of any common-mode signal is then

deferred until the differential signal has been adequately amplified. The final burden of CMR is then placed upon some circuit specifically chosen for this purpose.

Refer to Fig. 13-13 for a simple version of this concept. If R_i is the

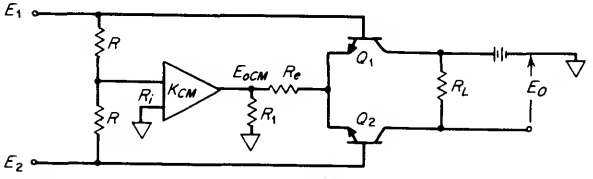


FIG 13-13 Simple differential amplifier with low CMR ratio.

input resistance of the common-mode amplifier with gain K_{CM}, the output of that amplifier is

$$E_{oCM} = \frac{K_{CM}}{1 + R/2R_i} \frac{E_1 + E_2}{2} \tag{13-6-1}$$

With only a common-mode signal at the input, $E_1 = E_2 = E_{CM}$ and

$$E_{oCM} = \frac{K_{CM}}{1 + R/2R_i} E_{CM} \tag{13-6-2}$$

but if a balanced differential signal is applied, $E_{oCM} = 0$. By adjusting K_{CM} we can make $E_{oCM} = E_{CM}$, and the common-mode signal produces no current in either Q_1 or Q_2. For the desired differential signal, Q_1 and Q_2 operate as a normal differential to single-ended amplifier.

Figure 13-14 shows a method for obtaining low balance and CMR ratios while also achieving gain accuracy with feedback. We start the discussion with amplifier A_2, which is a differential to single-ended amplifier. Let A_2 have a differential gain K_{2d}, a common-mode gain K_{2CM},

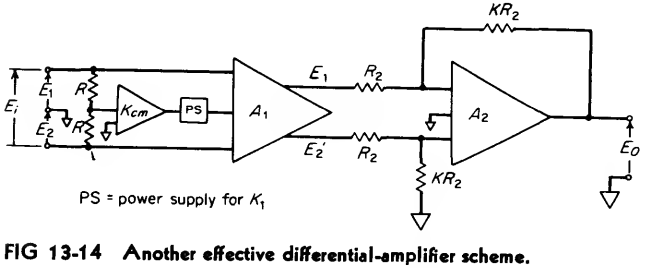


FIG 13-14 Another effective differential-amplifier scheme.

and arbitrarily high input resistances from each input to ground and across the two inputs.

With these assumptions an analysis of the circuit shows that

$$\frac{\partial E_0}{\partial E_1'} = \frac{(K/1 + K)(-K_{2d} + K_{2CM}/2)}{1 + (K_{2d} - K_{2CM}/2)/(1 + K)} \equiv F_1 \tag{13-6-3}$$

and

$$\frac{\partial E_0}{\partial E_2'} = \frac{(K/1 + K)(K_{2d} + K_{2CM}/2)}{1 + (K_{2d} - K_{2CM}/2)/(1 + K)} \equiv F_2 \tag{13-6-4}$$

It follows that

$$E_0 = F_1E_1' + F_2E_2'$$

$$= (F_1 + F_2)\,\underset{\substack{\text{common-}\\\text{mode}\\\text{gain}}}{}\,\underbrace{\frac{E_1' + E_2'}{2}}_{\substack{\text{common-}\\\text{mode}\\\text{signal}}} + \underbrace{\frac{F_1 - F_2}{2}}_{\substack{\text{differ-}\\\text{ential}\\\text{gain}}}\,\underbrace{(E_1' - E_2')}_{\substack{\text{differ-}\\\text{ential}\\\text{signal}}} \tag{13-6-5}$$

For the common-mode gain, we find

$$F_1 + F_2 = \frac{[K/(1 + K)]K_{2CM}}{1 + (K_{2d} - K_{2CM}/2)/(1 + K)} \approx \frac{KK_{2CM}}{K_{2d}} \tag{13-6-6}$$

under the normal conditions that $|K_d| \gg |K|$, $|K_{CM}|$. Under these same conditions the differential gain is

$$\frac{F_1 - F_2}{2} = \frac{-[K/(1 + K)]K_{2d}}{1 + (K_{2d} - K_{2CM}/2)/(1 + K)} \approx -K \tag{13-6-7}$$

so that from E_1' and E_2' to the output the

$$\text{CMR} = -\frac{K_{2CM}}{K_{2d}} \tag{13-6-8}$$

At this point let us note that the above analysis assumes that the two resistors are each set to exactly KR_2 in value. By adjusting one of these resistors, the CMR could be set to zero at low frequencies. A good design will concentrate upon minimizing K_{2CM}, however.

The first stage A_1 could be constructed from two amplifier blocks as shown in Fig. 13-15. If the voltage gains of the blocks are high, the differential gain of the stage is

$$K_{1d} = 1 + \frac{2R_1}{R_g} \tag{13-6-9}$$

The common-mode gain of the first stage is established by the common-mode amplifier to be very nearly unity since all ground returns in A_1 are

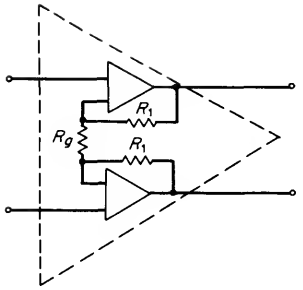


FIG 13-15 The first stage in Fig. 13-14.

referred to $(E_1 + E_2)/2$. Therefore, Eq. (13-6-6) gives the common-mode gain of the whole system, and the system CMR ratio is

$$\text{CMR} = \frac{K_{2CM}}{K^2(1 + K)(1 + 2R_1/R_g)} \tag{13-6-10}$$

Measurement of Gains and CMR Ratios. The common-mode gain of an amplifier is often measured directly by connecting the two input terminals of a differential amplifier together and applying a test signal between these terminals and ground. If the loading effects of a test voltmeter are small enough, the same voltmeter can be used to measure both common-mode input and differential output. However, since the output may be very small in some amplifiers, it may be desirable to use a selective voltmeter tuned to the frequency of the test signal. A wave analyzer is excellent for this application.

Ideally, and by definition, a balanced test signal and a differential voltmeter should be used to measure differential gain. The equipment required to follow this procedure may not be readily available, however, and usually no serious error occurs if one input terminal is grounded and a test voltage E_i is applied to the other terminal. This makes the common-mode input $E_{CM} = E_i/2$, but the common-mode output is still usually much lower than the differential output.

Some differential amplifying devices, such as electronic voltmeters, wave analyzers, and oscilloscopes, give their own output indications. The CMR ratios of these devices can be measured directly without a separate voltmeter; only a stable signal source is needed. The following procedure is used:

1. Apply a differential signal to the input, and adjust its amplitude to give a convenient output indication.

2. Short the two input terminals together, apply the signal source as a common-mode input, and adjust the input level to give the same output indication as in 1, above.

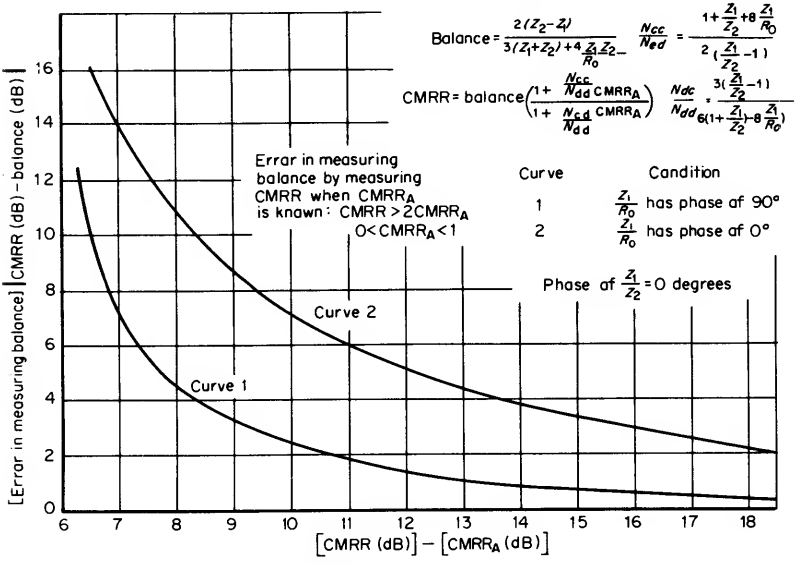


FIG 13-16 Error in measuring balance.

3. Connect the input in 2 to the device differentially, and measure its amplitude with the device being tested. The ratio of the input in 1 to the input in 2 is CMR.

If the device being tested is linear and has a range switch, a quicker procedure can be used. In step 2, do not change the test signal amplitude, but instead, simply change the setting of the range switch to get an on-scale reading at the output. The ratio of second reading to first is CMR ratio.

Balance is not an easy parameter to measure. Even though it is common for balance to be lower (because of high impedance to ground) than the CMR_A of the associated amplifier, the usefulness of the instrument depends upon the value of the greater of the two quantities. If CMR_A of the amplifier is significantly lower than the CMR of the total instrument, Eq. (13-4-16) reduces to CMR = CMR_B, and the balance can be measured directly. The accuracy with which this measure can be made is indicated on the curve in Fig. 13-16.

13-7 Dynamic Range and Distortion

Probably the most widely accepted meaning of dynamic range is the ratio of specified maximum signal level of a system or component to its

noise or resolution level, usually expressed in decibels [5]. While the maximum signal level is limited by a nonlinearity of the system, the noise or resolution level is not always measured in a uniform way. Even distortion at low levels can limit dynamic range, for a common requirement of communications equipment is the processing of both large and small signals through a common channel. In this situation the equipment must not generate spurious signals comparable in level to the smallest ones processed, and the equipment is normally evaluated by using a harmonic or intermodulation distortion measurement.

The *dynamic* range specification of a system input may be quite different from the unqualified range specification. For instance, a system with variable gain may have quite a large input range if its output noise level is measured with the highest gain setting and its maximum signal capability is measured with the lowest gain setting. The measurement of the noise or resolution level must be made with the system in a state that will handle the maximum signal level without overloading. When the dynamic range is limited by generation of spurious signals by the system, the maximum signal must actually be present during the measurement. Often the specified maximum signal level may not be close to the saturation level; that is, at such times the maximum level is limited by a small nonlinearity that causes spurious signals near the noise level.

There is no single best way to specify and measure dynamic range. The measurement technique generally simulates the worst condition that might occur when the system being tested is in operation. In a system which has a minimum signal level limited by phenomena other than spurious responses, the maximum signal specification may be checked by observing the output signal with an oscilloscope and looking for signs of clipping or other distortion which would cause unsatisfactory service. The signal used should be the worst condition expected in service, considering frequency, duty cycle, etc.

Measurement of the minimum signal level specified for a system is usually made by measuring either noise or distortion products, but may require other techniques if other factors limit operation at low signal levels. In measuring the noise level of a system, several conditions must be considered. When wideband noise is being measured, for instance, the bandwidth of the measuring instrument must include the frequency range of the system and in some cases must have a specified frequency response which is not necessarily flat. If the instrument responds to average, its reading will be about 1.1 dB lower than the true rms value of the noise level and will depend upon the characteristics of the noise. The noise level of a device or system is specified sometimes at a particular frequency and usually with a specified bandwidth. The specified bandwidth should be used in making the measurement, but as a compromise,

some similar bandwidth may be used and the measurement normalized to the specified bandwidth. The relationship that noise level is proportional to the square root of the bandwidth gives a valid conversion if the noise spectrum is flat over the widest bandwidth of interest and measurement of noise is specified to be made with a flat-response instrument.

Distortion. When a signal is processed, the output of a system may not be a faithful reproduction of its input. If the frequency response of the system is not flat, different frequency components of the input signal are amplified or attenuated by different amounts and the system will cause frequency distortion. Measurement of *frequency distortion* is usually made by measuring frequency response.

Phase distortion of a signal is caused by its different frequency components being delayed by unequal amounts of time. When all components are delayed equal amounts of time, such as in an ideal transmission line, their phases θ are delayed proportionally to their frequencies so that $d\theta/d\omega$ is constant. If the value of $d\theta/d\omega$ is dependent on frequency, different frequency components are delayed different times and the signal undergoes phase distortion. This can be seen by recalling that in the Fourier series representation of a waveform, the phase of each frequency component must be specified to define the waveform uniquely. While the correct phase relationship may be relatively unimportant in some signals, such as an acoustical waveform, other signals such as a train of data pulses may be completely destroyed by phase distortion, although the amplitudes of the component frequencies may be unaltered. See Chap. 2 for a basic treatment.

Measurement of phase distortion is generally expressed as the difference between the $d\theta/d\omega$ at two frequencies, one frequency being a reference. Many discrete measurements of $\Delta\theta/\Delta\omega$ or a swept measurement of $d\theta/d\omega$ gives phase distortion as a function of frequency over some range of interest.

Traditionally $d\theta/d\omega$ has not been measured directly, the accurate measurement of phase at different frequencies being difficult; instead, a method developed by Nyquist and Brand has been used [6]. This technique evaluates an AM signal which has been fed through the system to be tested. It can be shown that measurement of the phase of the modulation envelope yields values of $d\theta/d\omega$ equivalent to phase measurements made at the carrier frequencies and is much easier to accomplish accurately. Phase distortion measured by the Nyquist method is called *envelope-delay* or *group-delay* distortion and is generally done with an instrument especially made for this purpose.

A third distortion called *amplitude*, or *nonlinear*, distortion is the type most commonly dealt with and is caused by a nonlinear transfer characteristic. Amplitude distortion can be evaluated in a few cases, especially

if the system is dc coupled, by determination of the transfer characteristic and noting its departure from a straight line. This type of measurement is not generally made, however, since it lacks sensitivity, and results are not directly meaningful in most system applications. One frequent exception is in systems that are required to process amplitude levels accurately. Here a direct measurement of the nonlinearity is most meaningful.

A more common way of measuring amplitude distortion is by *harmonic* or by *intermodulation* distortion measurements. Both methods require driving the system with a signal at the specified maximum level and examining the system output for frequency components which were not present at the input, but were generated by the system nonlinearity. In making a harmonic-distortion measurement, a spectrally pure sinusoid is used as the driving signal. If such a signal is not available, it may be obtained by passing a sinusoid of limited purity through a low-pass filter which attenuates the harmonics a required amount. In any case, the requirement is that harmonic content of the driving signal be considerably below the level of the distortion level being measured. The harmonic content of the system output can then be determined by using a selective voltmeter or wave analyzer. Relative values of the various harmonics can be indicative of the nature of the phenomenon causing distortion. For example, push-pull systems tend to generate high odd-order products. Results are usually combined into a single specification of total harmonic distortion obtained by computing the square root of the sum of the squares of the harmonic amplitudes, that is, $(A_1^2 + A_2^2 + \cdots)^{1/2}$.

The ratio of this value to the fundamental amplitude may be expressed as a percentage or in decibels and is the value of total harmonic distortion normally specified.

Rather than use the wave analyzer to measure each harmonic individually, some prefer to use the distortion analyzer, which rejects the fundamental frequency and measures the average value of all harmonics simultaneously, the result being a single measurement of the total harmonic distortion. This method is less accurate but much quicker than the previous one. In making the measurement with the distortion analyzer it must be realized that components other than distortion products, such as hum and noise, may be present in the measurement results.

A second method of distortion measurement involves the passage of two large sinusoids through a system simultaneously and the measurement of the intermodulation frequency products at the output, with a wave analyzer. The sum of the amplitudes of the two input signals is generally the maximum signal level specified, but the frequencies and amplitude ratio can be chosen to be approximately a worst case. In

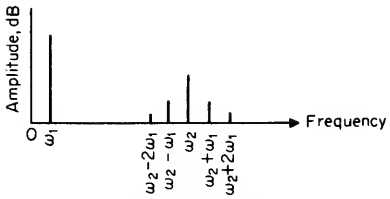


FIG 13-17 Intermodulation distortion by the SMPTE method.

standardizing this method, two conventions have become popular. The Society of Motion Picture and Television Engineers (SMPTE) has adopted a method in which one of the two sinusoids is 50 times the frequency of the other and one-fourth the amplitude. The frequency spectrum of the system output, shown in Fig. 13-17, is seen to consist of the two driving frequencies ω_1 and ω_2 and distortion products grouped around ω_2. The order of a distortion product is the sum of the coefficients of the two frequency terms; for example, $\omega_2 + 2\omega_1$ is odd distortion of the third order. In the SMPTE method, second-order distortion is defined as the ratio of the sum of the amplitudes of the two second-order components to the amplitude of the high-frequency signal ω_2. The third-order distortion is similarly defined, and the total distortion may be expressed as the square root of the sum of the squares of the second- and third-order terms and can be expressed as a percentage of the amplitude of the signal ω_2. An alternative method of measurement is to filter ω_1 from the output and express distortion as the modulation percentage of ω_2 found by envelope detection.

A second technique of intermodulation-distortion measurement introduced by the International Telegraph and Telephone Consultation Committee (Comité Consultatif International Télégraphique), the CCIT method, utilizes two closely spaced driving frequencies of equal amplitudes. The output of the system, including distortion products, is shown in Fig. 13-18. The even-order distortion can be taken as the ratio of the

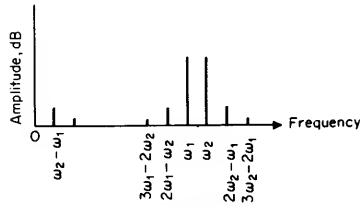


FIG 13-18 Intermodulation distortion by the CCIT method.

amplitude of the lowest-frequency component $\omega_2 - \omega_1$ to the sum of the output amplitudes of the two driving frequencies. Similarly, the odd-order distortion is taken as the ratio of the sum of the two third-order products to the sum of the amplitudes of the driving frequencies. The CCIT method is particularly useful for evaluating performance of a communications channel, where the driving signal and odd-order distortion products can be made to fall within a specified passband. Whether harmonic measurement or one of the intermodulation methods should be used depends mainly on the application and what kind of instrumentation is available. One advantage in making intermodulation distortion measurements is that the driving signals do not have to be spectrally pure since their harmonics, if moderate, will not cause significant intermodulation distortion and the harmonic frequencies themselves do not generally fall near the intermodulation products to be measured.

13-8 Slew Limiting

The small-signal frequency response curve of an amplifier gives its response to both high-frequency sinusoidal signals and transient pulses, provided that the maximum slew rate of the amplifier is not exceeded. The maximum slew rate, expressed in volts per second at the input terminals, or de_i/dt, is the time derivative of input voltage at which some stage in the amplifier is incapable of delivering enough current to charge the load capacitance of that stage in accordance with the equation

$$i = C\frac{de}{dt} \tag{13-8-1}$$

Theoretically, any stage in an amplifier could have the limiting slew rate, but often the first stage is the limiting one. This is especially liable to be the situation in solid-state amplifiers, because the first stage is frequently operated at a low value of quiescent current to reduce drift and random noise. When all the quiescent current is made to flow through a capacitor during a rapid change of voltage, the capacitor current cannot change further.

Of course, the load capacitance of an amplifier stage can be either stray capacitance or a lump capacitance. If shunt lump capacitance must be used to shape the frequency response of an amplifier, the designer should take slew limiting into consideration when deciding where to connect the capacitor. One design consideration is that the higher the open-loop gain is *after* a given stage, the lower the required de/dt is in that stage.

When an amplifier is driven by a sinusoidal signal $e_i = E \sin \omega t$, the

maximum de/dt occurs at values of ωt equal to 0 and π rad, or

$$\frac{de_i}{dt} = \omega E \cos \omega t \tag{13-8-2}$$

The maximum is proportional to both E and ω. Therefore, one way to measure maximum slew rate is to drive an amplifier sinusoidally at the maximum rated frequency and to increase the signal amplitude gradually until distortion begins to increase very rapidly. Knowing E and ω, one can then calculate de/dt by means of Eq. (13-8-2).

In a servorecorder, the maximum slew rate determines the maximum writing rate. A ramp waveform or a triangular wave of voltage can be applied to the input of a recorder amplifier, and the slope can be increased until the recorder no longer increases its writing rate. Similarly, a voltage step function can be applied and the maximum slew rate observed directly.

CITED REFERENCES

1. Linvill, John G., and James F. Gibbons: "Transistors and Active Circuits," McGraw-Hill Book Company, New York, 1961.
2. S-Parameter Techniques for Faster, More Accurate Network Design, *Hewlett-Packard J.*, vol. 18, no. 6, 1967.
3. Kuo, Benjamin C.: "Automatic Control Systems," p. 226, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962.
4. Kuo, Benjamin C.: "Automatic Control Systems," p. 236, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962.
5. Definitions of Terms for Analog Computers, *IEEE Stds. Publ. No.* 165, 1963.
6. Nyquist, H., and S. Brand: Measurement of Phase Distortion, *Bell System Tech. J.*, vol. 9, 1930.

CHAPTER FOURTEEN

MEASUREMENTS ON TRANSMITTERS AND RECEIVERS

D. B. Hallock

Collins Radio Company, Cedar Rapids, Iowa
with assistance of

W. B. Bruene

Collins Radio Company

The procedures for measuring and testing complete transmitting and receiving equipments often closely parallel those tests made on individual circuits or components. For example, measurement of the single-frequency power output of a complex high-frequency single-sideband transmitter consisting of many amplifying stages and internal frequency translations is made with the same instruments used for single-stage power measurements. The details of power measurement at various power levels and frequencies are discussed in other chapters. There are some measurements, however, that are unique to complete systems since they are dependent on the cumulative and interacting effects of many stages. It is the purpose of this chapter to describe such special measurements and to give some quantitative feeling for the performance of various classes of radio communication equipment.

It should be stressed that these special tests are quite subject to variation in technique as modified by the particular method of performance specification. As an example, the crossmodulation performance of a receiver may be measured with the undesired signal producing an audio output of 10 dB below reference level or 3 dB above the quieting audio level. The basic measurement technique remains the same.

14-1 General-performance Characteristics

Amplitude-modulated Systems. Amplitude-modulated (AM) receivers and transmitters are chiefly used for commercial broadcasting, radio navigation, and aircraft, maritime, and other mobile communications. Broadcast transmitters in the 535 to 1,605-KHz band are characterized by low audio-frequency distortion (5 percent or less), modestly wide audio bandwidth (100 to 5,000 Hz), low noise (50 dB below 100 percent modulation), and moderate frequency stability (± 20 Hz). Receivers for this service range widely in performance as a function of cost, but generally have low sensitivity (500 μV/m) and moderate audio distortion (2 to ten percent) at full power output. Commercial television broadcasting uses AM with a restricted-bandwidth lower sideband for the video signal. The transmission requirements include stringent adjacent-channel radiation level, linear phase and frequency response, and accurate pulse time bases. Quality television receivers have excellent sensitivity (5-dB noise figure), wide dynamic range of automatic gain control, and excellent pulse response. Amplitude-modulated equipment used for navigation purposes ranges from simple direction finders to rather complex aircraft landing systems. The chief characteristics of the latter are accurate modulation frequency and percentage in the ground station and low-distortion audio recovery in the receiver. Very high frequency transmission is used and receiver noise figures are in the range of 10 to 15 dB. Commercial-aircraft voice-communication equipment also operates in the very high frequency spectrum (118 to 132 MHz) and military voice communication occupies the 225 to 400 MHz region. These equipments are designed to provide reliable speech communication and have good sensitivity (9-dB noise figure), restricted audio bandwidth (300 to 3,500 Hz), speech processing such as clipping and compression, modest power output (25 W), and excellent receiver-overload performance. Maritime and other AM services in the high-frequency range (2 to 30 MHz) are also voice communication systems with restricted audio bandwidth. They are rapidly being replaced, however, with single-sideband equipment to conserve the rf spectrum.

Single-sideband Systems. Single-sideband equipment is chiefly used in the high-frequency range for voice and data communication. The chan-

nel bandwidth is usually 3 kHz, and as many as four channels are used, symmetrically placed about the suppressed carrier frequency. Transmitter power levels vary from as low as 5-W peak effective power for land mobile transceivers to several hundreds of kilowatts in large commercial installations. Receiver and transmitter third-order intermodulation distortion ranges from 25 dB below each tone of a two-tone test signal for simple voice equipment to 50 dB down for low-error-rate data transmission. Receiver sensitivity is modest (approximately 12-dB noise figure) since atmospheric noise usually exceeds the receiver internal noise sources. Differential time delay across a channel in single-sideband data systems must be held very low, 500 μsec of overall envelope delay being a typical specification.

Angle-modulated Systems. The term *angle modulation* is descriptive of FM (frequency modulation) and PM (phase modulation) and covers a wide range of radio equipment for voice, data, and entertainment uses. Very high frequency FM voice-communication equipment (police, taxicabs, etc.) uses 5- to 15-kHz deviation and has transmitter power output of from 30 W typically for a mobile transceiver to 100 W or more for a base station. Both transmitter and receiver spurious output and responses are low (85 dB down or more) and receiver noise figure is low, in the order of 5 to 10 dB. Speech clipping and other voice processing are often used. The mobile equipment is very ruggedly constructed and is often completely solid state, including the transmitter output amplifier. Phase data-communication equipment is chiefly used by the military in the 225- to 400-MHz range and commonly has biphase modulation to transmit binary data. Entertainment FM broadcast operates in the 88- to 108-MHz band in 200-kHz channels with ± 75-kHz peak deviation. The transmitter audio distortion is very low (3.5 percent maximum) and background noise is greater than 60 dB below peak deviation. Frequency-modulation transmitters for stereophonic broadcasting use a subcarrier system which requires excellent phase linearity to achieve the 29.7-dB channel-separation requirement in the United States. Frequency-modulation broadcast receivers vary widely in performance as a function of price; the better receivers have noise figures as low as 5 dB, excellent quieting, and at least 30-dB channel separation for the stereophonic units.

The aural channel of television transmission also uses FM with characteristics similar to those for FM broadcasting except that the maximum deviation is ± 25 kHz and FM noise is required to be at least 55 dB below peak deviation. The aural carrier is positioned 4.5 MHz above the visual carrier, and the receiver recovers the audio with a discriminator tuned to the 4.5-MHz beat between aural and visual carriers.

The preceding paragraphs have briefly outlined some of the chief oper-

ational characteristics of radio communication equipment in order to give a basis for the level of performance expected when measurements are made on a complete unit or system. A more complete listing of the specific commercial and government publications relating to radio equipment specifications is given at the end of this chapter.

14-2 Basic Measurements

It has been noted that the measurement techniques in other chapters also apply to complex receiving and transmitting equipment. For receivers, these common tests include sensitivity, noise figure, crossmodulation, intermodulation, selectivity, frequency stability, spurious responses, distortion, and audio power output. Common transmitter measurements are power output, harmonic content, distortion, intermodulation, frequency stability, spurious outputs, and overall efficiency.

In applying the basic methods to complete equipments, the practical limitations of both the test equipment and the test item must be understood. A good example of this is the crossmodulation measurement of an AM superheterodyne receiver. When the undesired test signal is spaced close to the desired-signal frequency, it is often observed that the receiver's audio output is noise rather than a tone at the undesired-signal modulation frequency. This effect is caused by spectral noise of either the undesired-signal generator or the receiver's local oscillator, or both. For instance, if the noise power spectral density of the undesired-signal generator is high at the desired-signal frequency, that noise will be heterodyned directly to the intermediate frequency and appears as noise in the detected audio output. The converse is true if the local-oscillator noise level is high at a frequency removed from the undesired signal by an amount equal to the intermediate frequency.

14-3 Special System Measurements

There are a large number of specialized tests on transmitting and receiving equipment and systems needed to characterize the overall performance. Only the more common types are discussed in the following paragraphs since the detailed equipment specifications and design often dictate measurement techniques peculiar to a given case.

Conducted and Radiated Interference. Radio equipment generally contains internal sources of rf energy which may be either conducted out of the equipment by the power, control, or signal cables or radiated directly from the equipment enclosure. For instance, the local oscillator in a superheterodyne receiver will cause a small amount of voltage to be present at the antenna input terminals. In a multiple-receiver installation,

one receiver could easily be tuned to the local-oscillator frequency of another receiver, and this causes an interference problem.

Conducted Interference. A test setup for measuring the conducted interference present on power or control leads is shown in Figure 14-1. For reproducible results, the test should be carried out within a shielded room or chamber and the equipment mounted on a large, grounded sheet of brass or copper. The equipment under test should be bonded to the grounded sheet in the same manner as it would be in an actual operating installation. The current probe is a toroidal transformer through which each control or power lead is passed one at a time. The calibrated receiver is a wide-range superheterodyne which contains means for accurately measuring the voltage level appearing at its input terminals. Such receivers are called *electromagnetic-interference (EMI) meters* and are manufactured by such companies as Stoddard Electro Systems. Alternatively, a series of uncalibrated receivers may be used together with signal generators to measure the amplitude of each narrow-band emission by the signal-substitution method. The frequency-range requirements for the test depend on the specification requirements; MIL-STD-461, for example, requires measurements from 20 Hz to 400 MHz. The test limits of MIL-STD-461 in the frequency range of 20 kHz to 50 MHz are shown in Fig. 14-2. For receivers, the conducted-interference test should be performed both with no signal input and with a large signal input which results in maximum rf levels within the receiver.

Measurement of the conducted energy appearing at the antenna terminal of a receiver or transmitter is effected with the test shown in Fig. 14-3. When a transmitter is being tested, the attenuator is actually the output load for the equipment and must be capable of dissipating the rated output power. The rejection network is a bandstop filter and serves to increase the dynamic range of the calibrated receiver or EMI meter by attenuating the fundamental frequency. This filter can be a

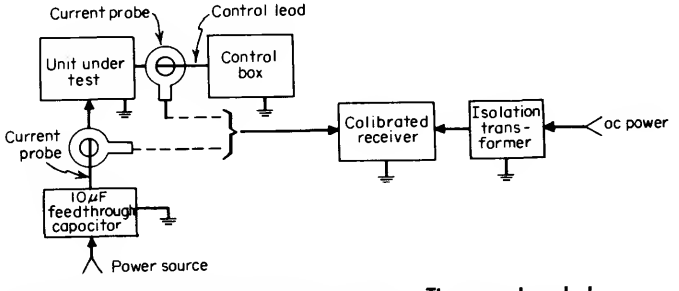


FIG 14-1 Conducted-interference test setup. The ground symbol indicates a bond to a common ground plane.

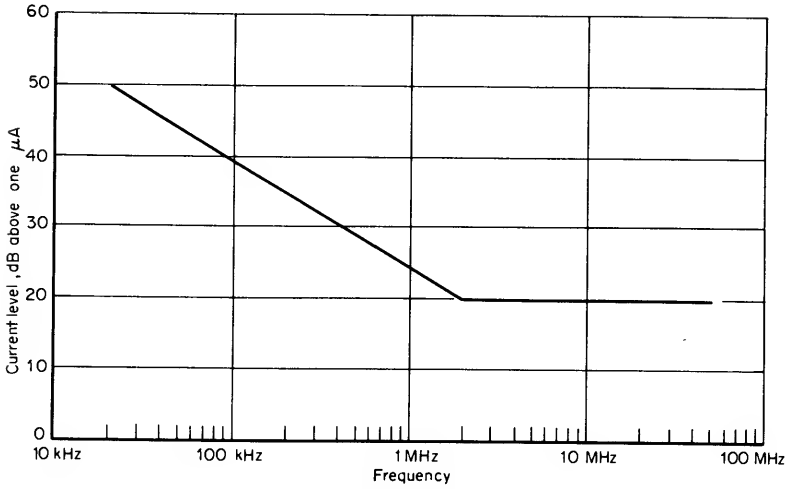


FIG 14-2 Conducted-interference test limits for power and control leads. (From MIL-STD-461.)

simple parallel-tuned *LC* network in a shielded box when the measurement frequency is well removed from the transmitter output frequency. For close frequency spacings, a quartz-crystal bandstop lattice filter or a phase-locked cancellation loop may be required. The rejection network is not required when measuring the conducted interference of a receiver or of a transmitter under key-up conditions. The term *key-up* indicates the state in which all circuits in the transmitter are energized except those that control the actual emission of desired output. When the rejection network is used, its frequency response must be known or measured to allow calculation of the power actually flowing at the antenna terminals. There are two possible sources of spurious responses in this measure-

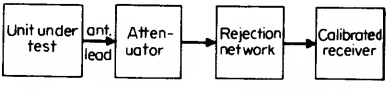


FIG 14-3 Antenna conducted-interference test setup.

ment that must be distinguished from actual antenna-conducted output. These sources are nonlinearity in either the rejection network or the EMI meter. If the power level is sufficiently high, the rejection network can generate harmonic energy because of core saturation in powdered

iron- or ferrite-cored inductors, if used. The EMI meter is a superheterodyne receiver with its own spurious responses, notably the image frequency. Network or high-order receiver spurious products can be identified by increasing the attenuator setting by a small amount (3 dB) and noting the indicated level decrease on the EMI meter. If it is the same, then a true output is being observed. If the EMI-meter decrease is larger than the attenuation increment, then a nonlinearity is at least partially responsible for the indicated reading. The present state of the art allows approximately 1 mW as the maximum available power supplied to the EMI meter input for spurious free readings. Image responses may be identified by linearly combining the output of a stable signal generator with the signal output of the rejection network and applying the sum to the EMI meter. The signal generator frequency is adjusted to produce a zero beat with the spurious signal. The tuning dial of the EMI meter is then moved a small amount; if the zero-beat note is unchanged, the meter response is due to conducted interference; if the beat note varies in frequency, then the indication is due to image or mixer spurious responses in the EMI meter.

As an example of the test limits for this measurement, Fig. 14-4 shows the harmonic and spurious emission requirements of MIL-STD-461. The measured frequency range depends on the test specification but can be as wide as from 10 kHz to 40 GHz.

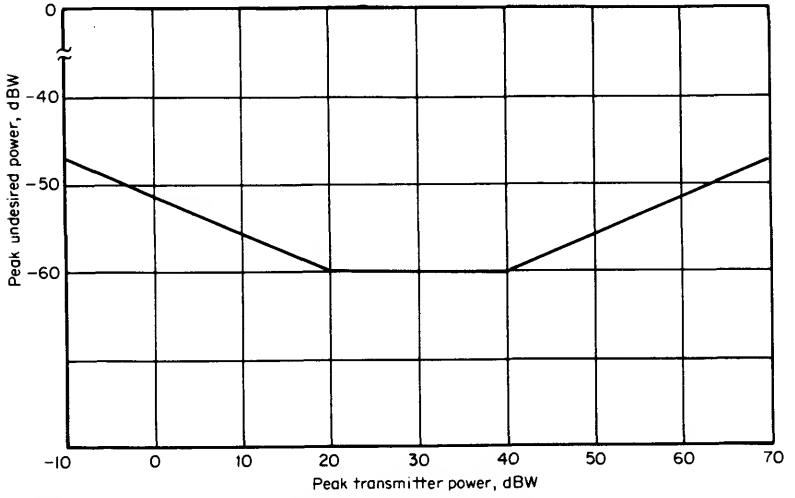


FIG 14-4 Harmonic and spurious limits for antenna conducted interference. (From MIL-STD-461.)

Radiated Interference. The emission of unwanted energy from equipment can be in the form of a nearby magnetic field or an electromagnetic field escaping through seams in the enclosure or other openings. The magnetic-field measurement is carried out with a loop probe, while the electromagnetic field emission is sensed with antennas having known effective height at the test frequencies involved.

The magnetic-field measurement is conducted by positioning the loop probe a given distance (typically 7 cm) from one face of the equipment, applying the output voltage of the loop to a calibrated receiver, and scanning a specified frequency range for points of peak emission. The loop voltage at each peak point is then converted to the magnetic-flux-density units called for in the test specification. Again referring to MIL-STD-461, the unit of flux density is decibels above 1 picotesla (dBpT). The induced loop voltage is related to flux density by

$$E = \pi \sqrt{2} f A N B \times 10^{-4}$$

where E = rms-induced voltage, V
f = frequency, Hz
A = loop area, cm^2
N = number of turns in the loop
B = flux density, T

The standard loop has 36 turns of wire and has an area of 139 cm^2. The frequency range is specified from 30 Hz to 30 kHz, and the test limits are 140 and 20 dBpT, respectively. These limits result in an induced voltage of 666 μV at 30 Hz and 0.666 μV at 30 kHz.

Determination of the radiated electromagnetic field is carried out in a shielded room having dimensions such that the receiving antenna is at least 1 m from any obstruction and is positioned 1 m from the equipment under test. The walls of the room should be covered with carbon-filled rf absorption material to suppress reflection of electromagnetic energy. The receiving antenna is connected to a calibrated receiver such that a value of field strength in microvolts per meter can be obtained. The range of frequencies tuned by the receiver should at least cover the operating frequency range of the equipment. The specification in MIL-STD-462 requires measurements from 14 kHz to as high as 10 GHz.

Electromagnetic Susceptibility. Radio equipment is expected to operate in a hostile environment from an electromagnetic viewpoint, especially when large systems with several transmitters, receivers, and antennas are simultaneously operating at one site. Quantitative measurements are needed of performance degradation as a function of unwanted signals conducted through power and control wires into the equipment or radiated into the enclosure from external sources.

Conducted Susceptibility. In the 30-Hz to 50-kHz frequency range, susceptibility to undesired signals on the input power leads (or control leads if required) is measured by inserting the undesired signal in series with the lead and using a transformer with low leakage inductance (1 μH

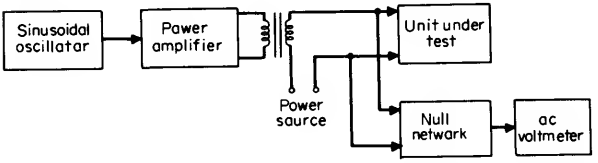


FIG 14-5 Conducted-power-lead susceptibility measurement (30 Hz to 50 kHz).

or less). The test setup is shown in Fig. 14-5 for injecting the interference into the equipment power lines. The output impedance of the power amplifier measured at the transformer secondary must be low enough to cause only a small voltage drop in the power supplied to the unit under test. The null network is required to reject the power supply frequency (if ac power is used) to simplify measurement of the undesired signal by the ac electronic voltmeter. The measurement procedure consists in applying a specified undesired voltage level to the equipment under test and then measuring a critical performance parameter of the equipment. For a receiver, this might be a sensitivity or noise-figure measurement; for a transmitter, it might be the carrier signal-to-noise ratio.

To measure conducted susceptibility at higher frequencies (typically 50 kHz to 400 MHz) the test setup in Fig. 14-6 is used. Several signal generators and power amplifiers are usually necessary to cover the frequency range. Also, the coupling capacitor has to be changed over the frequency range, with the use of types having a total reactance of 5 Ω or less at the frequency of measurement. Physically short leads must be used in the test to ensure that the rf voltmeter is accurately measuring the voltage at the power input terminals of the unit under test. As in

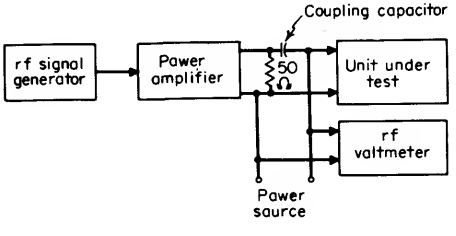


FIG 14-6 Conducted-power-lead susceptibility measurement (50 kHz to 400 MHz).

the low-frequency susceptibility test, a specified rf voltage adjustable over a band of frequencies is applied to the power line while measuring key performance parameters of the equipment.

14-4 Measurements on Receiving Systems

There are several measurements applying to complete receivers which differ from the basic tests and yet are regularly performed. The following sections discuss these measurements.

Quieting Characteristic. The output signal of a receiver consists chiefly of random noise when no signal is being applied to the antenna terminals. This noise is band limited by the audio response of the receiver, and, in the case of superheterodyne receivers, is often band limited by the intermediate-frequency selectivity. The noise power comes from the amplified thermal (Johnson) noise of the receiver input-terminal source resistance (the resistance that the receiver input "sees") and from the noise sources within the receiver itself. When the receiver is connected to an antenna, additional input noise is introduced from galactic, atmospheric, and man-made sources.

When an unmodulated signal is applied to the receiver, the noise output generally decreases. In AM and single-sideband receivers, this is due to the reduction in gain as the automatic-gain-control circuit acts to hold the detector signal level constant for variations in received signal level. In FM receivers, the noise output drops as the input signal captures the limiting stages in the receiver. This property of a receiver in the presence of an unmodulated input signal is termed the *quieting characteristic.*

The quieting characteristic is not generally a linear function of the input signal. For very small signals, typically 1 μV or less, the receiver is operating in the threshold region. In this region, AM receivers exhibit first a small increase in noise output and then a monotonic decrease as the signal level is increased from zero. The noise increase with a very weak signal is caused by an increase in the envelope-detector efficiency with the presence of a sinusoidal signal of amplitude comparable with the rms noise amplitude. The sinusoid acts as a carrier with which the noise components can heterodyne to produce an increase in detected noise. At higher signal levels, the automatic gain control acts as noted to reduce the noise output. In single-sideband receivers, however, the noise output does not vary for signals below automatic-gain-control threshold where the gain of the receiver is constant. This is so for a well-designed receiver which linearly translates an rf signal down to an audio frequency. The small signal produces an audio tone at the receiver output; this increase in output is to be distinguished, however, from the noise output. In FM receivers, the character of the noise output changes in the thresh-

old region. With no signal input, the noise power spectral density is substantially flat within the audio bandwidth if no frequency-shaping deemphasis networks are used. As the signal is increased, the subjective effect is that the flat noise tends to decrease but is replaced by a "popping" sound. This happens as the desired signal starts to saturate the limiting stages in a receiver. Noise peaks, however, momentarily saturate the limiters, which produces a rapid phase change in the signal to the discriminator stage. The discriminator detects the phase change and produces a pop in the audio output. As the carrier-to-noise ratio in the receiver prior to limiting increases over 0 dB, the popping sound rapidly diminishes in amplitude and the receiver quiets.

In all types of receivers with a strong input signal, the amount of quieting reaches a plateau region where an increase in signal produces little reduction in noise output. In a well-designed receiver the limitation is chiefly in the noise and hum in the audio stages; typical maximum quieting ratios are 40 to 60 dB below the noise with no signal present. Amplitude-modulation and single-sideband quieting characteristics commonly exhibit changes in slope as the stages in the receiver automatically change gain at different rates as a function of input signal.

Measurement of the quieting characteristics is applicable to AM, single-sideband, and FM receivers and is indicative of the automatic gain control and the gain distribution of AM and single-sideband equipment and the limiter performance in FM equipment. The measurement is simply made by applying an unmodulated signal input to the receiver and measuring the total noise power output of the receiver as a function of signal input level. The amount of quieting is usually expressed in decibels with respect to some reference: For AM receivers, the reference might be the audio output with a 1-mV rf input signal modulated 30 percent at 1,000 Hz; for FM receivers, the reference is the noise power output with no signal input; for single-sideband receivers, the reference can be the audio output with a 1-mV unmodulated rf signal whose frequency is set to the middle of the channel. This measurement for single-sideband receivers is complicated somewhat by the output audio tone present with a single-rf input. One method of measurement is to remove the tone with a null network or total-harmonic-distortion analyzer. Another method is to note the receiver's automatic-gain-control line voltage in the presence of the test signal, remove the test signal, use a battery or other source to make the automatic-gain-control voltage artificially equal to that with signal, and then measure the noise power output.

Impulsive-noise Rejection. Fixed and mobile voice-communication receivers commonly employ squelch circuits, which mute the audio output when a signal is not present. It is functionally desirable that the receiver remain muted or squelched off in the presence of strong impulsive

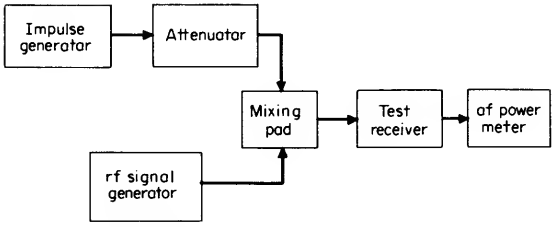


FIG 14-7 Impulsive-noise rejection test setup.

noise such as ignition pulses. Quantitative measurements may be made with the setup shown in Figure 14-7. The impulse generator must be calibrated in some manner, such as the noise power spectral density in decibels above 1 μV/MHz (dB-μV/MHz). An excellent wideband (1,000 MHz) impulsive noise source is the type having a 60-Hz mechanical relay with mercury-wetted contacts. The noise-rejection measurement is made in two steps. First, the rf-signal generator is reduced to zero output, and the noise power spectral density required to open the squelch is determined. (The squelch sensitivity must initially be set to the equipment specifications.) Second, the rf signal generator (unmodulated) is set to some specified level below squelch sensitivity (typically two-thirds of the rf input squelch sensitivity), and the noise power spectral density to open the squelch is again determined. Military test limits (MIL-STD-461) for these two measurements are 90 and 50 dB-μV/MHz, respectively, for not opening the squelch.

Many squelch circuits are designed to open at a predetermined signal-to-noise ratio, typically in the range of 1 to 3 dB. Measurement of the ability of such a circuit to open only in the presence of a usable signal-to-noise ratio is made by applying a given noise power spectral density and then increasing the rf signal until the squelch opens. The noise generator is then turned off and the drop (in decibels) in receiver power output is noted. This test is repeated for several values of noise level to characterize the squelch performance. The drop in receiver power output measured in this way is actually the ratio of the signal plus noise to the noise and can be converted to signal-to-noise ratio by noting that

$$\frac{S}{N} = \frac{S+N}{N} - 1$$

where S and N are numerical signal and noise powers, respectively.

The mixing pad shown in Fig. 14-7 consists simply of three resistors connected in a Y configuration. For 50-Ω systems, the resistors are all equal to 16.7 Ω (18-Ω carbon-composition resistors are a practical com-

promise). When connected to 50-Ω sources and load, the resistance looking into any port is 50 Ω and the attenuation between any two ports is 6 dB.

14-5 Sinad Sensitivity

This type of receiver sensitivity test is used primarily with voice-communication equipment since it gives a measurement that is more indicative of the expected intelligibility for a given signal strength than the usual ratio of the signal plus noise to the noise. *Sinad* is an acronym of "signal plus noise and distortion." Although this test is commonly specified for FM equipment, it is equally applicable to AM (dual sideband) and single-sideband. The measurement technique consists in applying a signal of known amplitude and modulation to the receiver antenna input. The audio power output is then measured; this power consists of the total sum of the recovered modulation, the distortion products produced by receiver nonlinearity, and the amplified thermal noise of the signal source and receiver. A narrow-band-rejection filter, tuned to the modulation frequency, is then interposed between the receiver audio output and the power meter. A second power reading is obtained which now is due to the distortion-plus-noise output. The ratio of the first power to the second, expressed in decibels, is called the *dB sinad* for the given signal level and modulation.

Figure 14-8 illustrates the basic equipment required to perform the sinad test. Design-value impedance should be presented to the input and output of the receiver under test. The attenuation of the null network at the modulation frequency should be approximately 20 dB greater than the amount of dB sinad measured for negligible error due to audio feedthrough. Additionally, the rejection band should be narrow compared with the audio bandwidth of the receiver, since it is desired that

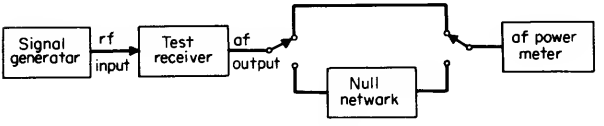


FIG 14-8 Sinad sensitivity test setup.

only the modulation frequency be rejected. The flat loss of the null network must be taken into account also.

Commercially available total-harmonic-distortion analyzers may be used to measure the audio output. These instruments typically contain a Wien-bridge null network together with the necessary switching and

metering. The output meter of these instruments is usually not a true rms (power) meter; the error is negligible with the meter damping that is provided. The null depth is typically in excess of 80 dB, allowing sinad measurement down to at least 60 dB. This is very adequate since the harmonic distortion of the receiver will not usually permit sinad ratios as high as 60 dB.

14-6 Modulation-acceptance Bandwidth

In FM voice receivers a measure is needed of the maximum deviation that can be accepted for a certain degradation in sinad sensitivity. This measure is called *modulation-acceptance bandwidth* and is characterized by the frequency deviation required to produce 12-dB sinad for a signal that is 6 dB larger than a signal with a standard modulation that produces 12-dB sinad sensitivity.

The measurement is made by applying a signal that is frequency modulated at 1,000 Hz and has a deviation that is two-thirds of the rated system deviation. The signal level is initially set to 1,000 μV, and the receiver audio gain control is adjusted for rated power output. Next, the signal level is reduced until a sensitivity of 12-dB sinad is obtained. Finally, the signal level is increased 6 dB over that just obtained for 12-dB sinad, and the deviation is increased until a sensitivity of 12-dB sinad is again reached. The increased deviation, multiplied by 2, is a measure of the modulation-acceptance bandwidth.

It can be seen that this test includes all sources of distortion in the receiver such as sideband attenuation in the intermediate-frequency filter, nonlinearity of the discriminator, and audio distortion. It therefore relates to speech intelligibility under conditions of transmitter frequency deviation at or exceeding the maximum design value for the system. Modulation-acceptance bandwidth is a dynamic measurement of performance, as contrasted with a normal measurement of selectivity, that uses a single frequency which is slowly varied through the receiver's passband.

14-7 Correlation of Sensitivity with Noise Figure

The techniques of measuring AM and single-sideband sensitivity of individual amplifiers and mixers at rf, as discussed in Chap. 16, also apply to complete receiving systems. Similarly, noise figure determination of a receiver embodies the principles given in Chap. 4. Since many equipment specifications give one form of sensitivity requirement, but not both, it is necessary to relate the two for design purposes.

Figure 14-9 shows an equivalent input circuit for a receiving system in

which the resistor R_n represents the room-temperature resistor that, when connected across the input terminals of a perfect receiver, will produce the same receiver noise power output as the real receiver does with a short-circuited input. The resistance R_s is the antenna source resistance

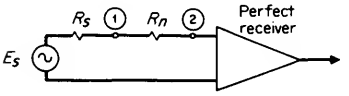


FIG 14-9 Noise model of receiver input.

at room temperature and is in series with an open-circuit voltage E_s. The noise factor F of the network consisting of R_s and R_n follows from the basic definition of noise factor and is

$$F = \frac{(S/N)_{\text{in}}}{(S/N)_{\text{out}}} = \frac{(S/N)_{\text{point 1}}}{(S/N)_{\text{point 2}}} \tag{14-7-1}$$

where the signal-to-noise ratios S/N are available power ratios. The available signal power at point 1 is $E_s^2/4R_s$, while the available noise power is KTB, where K is Boltzmann's constant, T is the temperature in degrees Kelvin, and B is the bandwidth of the receiver. Similarly, at point 2, S is $E_s^2/[4(R_s + R_n)]$, and N is still just KTB. With these relations, Eq. (14-7-1) reduces to

$$\begin{aligned} F &= 1 + \frac{R_n}{R_s} \\ FR_s &= R_s + R_n \end{aligned} \tag{14-7-2}$$

The total open-circuit noise voltage at the perfect receiver input (point 2), is

$$\bar{E}_n = \sqrt{4KTB(R_s + R_n)} \tag{14-7-3}$$

and, from Eq. (14-7-2),

$$\bar{E}_n = \sqrt{4KTBFR_s} \tag{14-7-4}$$

Considering now an AM system, the ratio of signal plus noise to noise power is

$$\frac{S + N}{N} = \frac{(ME_s)^2 + E_n^2}{E_n^2} \tag{14-7-5}$$

where M is the modulation index. This may be solved for E_n and equated

to Eq. (14-7-4) to give the general expression

$$F = \frac{M^2 E_s^2}{4KTBR_s[(S + N/N) - 1]} \tag{14-7-6}$$

The noise figure in decibels is just 10 log F. For the commonly encountered specific case where

$$\begin{aligned} M &= 0.3 \text{ (30\% amplitude modulation)} \\ K &= 1.38 \times 10^{-23} \text{ (Boltzmann's constant)} \\ T &= 300°\text{K} \\ R_s &= 50\ \Omega \\ \frac{S+N}{N} &= 10 \text{ (10-dB power ratio)} \\ B &= 2\text{AFBW (AFBW = postdetection noise bandwidth, Hz)} \\ E_s &= \text{voltage of source (open-circuit), } \mu\text{V} \end{aligned}$$

Equation (14-7-6) reduces to

$$F = \frac{6030 E_s^2}{AFBW} \tag{14-7-7}$$

Equations (14-7-6) and (14-7-7) should be applied with the understanding that the predetection bandwidth should be at least twice the postdetection bandwidth but not more than perhaps 10 times the postdetection bandwidth. Gannaway [1] and Fubini and Johnson [2] give more detailed discussion of the detection process as a function of a pre- and postdetection bandwidths.

For single-sideband receivers where the $(S + N)/N$ ratio has been determined for a given single-frequency input, Eq. (14-7-6) may be used with $M = 1$ and B equal to the overall equipment noise bandwidth. In this case, B may be determined primarily by intermediate-frequency selectivity rather than the audio stages following the product detector.

It is worthy of note that measurement of a receiver's noise bandwidth may be made more easily by a determination of sensitivity and noise figure than by measuring the steady-state response directly, squaring it, and finding the equivalent square-shaped bandwidth having the same area. For the rather special case of maximally flat (Butterworth) steady-state response, it is possible to find the noise bandwidth analytically by knowing the number of poles, n. The squared transfer function for Butterworth response [3] is

$$|G_{12}|^2 = \frac{1}{1 + x^{2n}} \tag{14-7-8}$$

where x is a normalized frequency variable equal to ω/ω_{3dB} for low-pass

TABLE 14-1 Noise Bandwidth of the Butterworth Response Shape in Terms of the Attenuation

n	α, dB
1	5.4
2	3.8
3	3.65
4	3.46
5	3.38

or QBW/f_0 for bandpass filters. Let W be the bandwidth of a response with square shape and amplitude 1 and with an area equal to that of the G_{12} response.

$$W = \int_0^\infty |G_{12}|^2 \, dx = \int_0^\infty \frac{dx}{1 + x^{2n}} = \frac{\pi}{2n \sin \pi/2n} \tag{14-7-9}$$

The response of the Butterworth shape at bandwidth W is

$$|G_{12}|^2 = \frac{1}{1 + W^{2n}} \tag{14-7-10}$$

when ω_{3dB} is normalized to unity. The attenuation in decibels is

$$\alpha = 10 \log (1 + W^{2n})$$

By using the above equations, the values in Table 14-1 are obtained. The bandwidth at attenuation points α is equal to the noise bandwidth. For a single tuned circuit $n = 1$, the noise bandwidth is equal to the bandwidth between points whose attenuation is 5.4 dB.

14-8 Automatic-gain-control Characteristics

Automatic-gain-control Rise. The degree of output-level control provided by a receiver's automatic gain control is simply defined by measuring the rise in audio output produced by a given increase in signal input level. This test is mainly applicable to AM and single-sideband receivers. Although the signal level and performance details depend on the particular use of the receiver, a representative test would be to apply a 1,000 μV rf input, modulated 30 percent by 1,000 Hz, to an AM receiver. The volume control is then adjusted for a rated audio power output. The signal level is dropped to 5 μV and the audio output power noted. The signal is then increased to 200,000 μV, and the audio output power is again measured. The increase in audio power output between the 5-

and 200,000-μV input-signal levels is called the *automatic-gain-control rise* and is usually expressed in decibels.

In making this measurement, it is well also to perform a total-harmonic-distortion measurement, particularly at the higher signal levels. This is

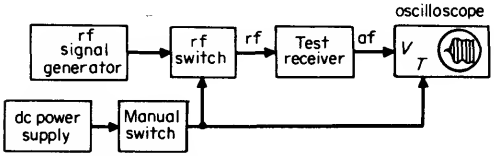


FIG 14-10 Automatic-gain-control settling-time test setup.

done since it detects audio or rf envelope clipping at any point in the receiver. The rf signal level should also be slowly varied over the specified range while any discontinuities or peaks of audio output are noted.

Settling Time. The manner in which a receiver output varies after application of a step in signal input is of interest since it characterizes the envelope distortion of AM and single-sideband voice receivers and the accuracy of special AM receivers used for navigation. Measurement of step response is made with the equipment shown in Fig. 14-10, which applies a modulated-rf step input to the receiver and simultaneously triggers the horizontal sweep of an oscilloscope. The receiver audio output is applied to the vertical amplifier of the oscilloscope. The sweep speed is adjusted so that the vertical deflection has stabilized at the end of the horizontal sweep. A usual definition of settling time is the time required for the output to arrive at 90 percent of its steady-state value, including overshoots. For detailed study of the response shape, a storage oscilloscope or a camera may be used to retain a single sweep image.

The isolation provided by the rf switch in the off state is of particular concern with sensitive receivers. For example, if a 0.1-V step input is to be applied to a receiver having an automatic-gain-control threshold of 1 μV, then an attenuation of 100 dB is required in the switch for the initial condition to be essentially zero signal. A mechanical coaxial relay with internal shielding and an output grounding contact will usually suffice for slowly responding receiver automatic gain control. For fast responses, it may be necessary to use cascaded diode-ring modulators with a dc step control applied to the port having response down to zero. From the standpoint of attenuation, keying the oscillator of the signal generator off and on is ideal. This usually requires modification of the generator

and, for narrow-band receivers, is unsatisfactory since a frequency transient (chirp) occurs during the transition interval.

Countermodulation. In an AM receiver, particularly the kind used for aircraft navigation, a phase shift of the modulating signal occurs in the receiver because a small amount of recovered audio appears on the automatic-gain-control line, which changes the receiver gain slightly at the modulation rate. This effect, called *countermodulation,* is the greatest at low modulation frequencies. The phase change is measured by comparing the modulation signal of the rf-signal generator with the audio output of the receiver as a function of rf signal level and modulation frequency. The phase shift of the signal generator modulation should be checked preliminarily by substituting an envelope detector of known linearity for the test receiver. The phase comparison can be made either with an oscilloscope as discussed in Chap. 11 or with an audio phase-shift meter. The amount of phase shift to be measured depends on the receiver application; a very high frequency navigation receiver for omnirange use might be limited to one degree of error at a modulation frequency of 30 Hz.

14-9 Measurements on Transmitting Systems

There are many specialized measurements taken on transmitters, determined by the particular equipment mode of operation and its specifications. This section will discuss some of the transmitting measurements applicable to systems and not considered basic device tests.

Single-sideband Distortion. A basic method of measuring the linearity of a single-sideband transmitter is to apply a two-tone test signal to the input and to observe the output spectrum on a spectrum analyzer. This has been discussed in Chap. 2.

A newer method of measuring the distortion produced by single-sideband transmitter nonlinearity is called the *noise-loading test.* This test has been necessitated by the extensive use at high frequency of multichannel single-sideband data transmission. For example, a 3-kHz voice channel may be used to multiplex by frequency division up to 18 uncorrelated frequency- or phase-shift-keyed signals. When the signals within a channel are uncorrelated, and channels are likewise uncorrelated to each other, the rf-envelope signal in the rf stages of the transmitter approaches noise with a Rayleigh amplitude distribution (central-limit theorem). Since the distortion products of all the other channels appearing in one channel tend to approach a noiselike signal, the error rate of the data system is affected by the drop in apparent signal-to-noise ratio. The noise-loading test provides a quantitative means of predicting digital data error performance when the results of the test are known.

The concept of the noise-loading test is rather simple. A white-noise signal (constant noise power spectral density over the frequency band of the transmitter) is applied to the input terminals and increased in amplitude until the transmitter power output is a reference value called *rated noise power*. This power level is sometimes defined as 7 dB less than the peak envelope power rating of the transmitter. A narrow spectral notch (approximately 10 percent of the system bandwidth) is then placed in the input noise, the input is increased slightly to produce rated noise power, and the output power spectral density lying in the notch is measured. The ratio of the power in the same bandwidth with the notch removed to that with the notch present is called the *single-sideband noise power ratio* and is usually expressed in decibels. Figure 14-11 shows the basic setup needed to perform the test. The bandwidth of the filter in the output of the transmitter under test should be small enough compared with the notch bandwidth so that only the power in the depth of the notch resulting from distortion products enters the power meter. An extension of Fig. 14-11 would be to replace the transmitter under test with a complete communications link including a transmitter, the propagation path, and the receiving system. The results of this test include the distortion products of both the transmitter and receiver, and the additive noise, as well, of the propagation path and the frequency sources (local oscillators) within the transmitter and receiver.

A practical setup for transmitter noise-loading distortion is shown in Fig. 14-12. This method differs somewhat in that a continuous display of noise power ratio (single-sideband) is obtained by simultaneously measuring output noise both in the notch band and in a signal band spaced close to the notch band. The ratio of the two noise powers is obtained logarithmically and presented continuously on a dc meter. The notch filter has a stop band centered at 1,600 $\pm$ 100 Hz and has a 3-dB bandwidth not exceeding approximately 10 percent of the channel bandwidth (250 Hz is a typical bandwidth). The notch-band skirts must be steep enough to present at least 50-dB attenuation to noise components at the passband frequency of the distortion filter. The distortion filter is a bandpass filter having a 3-dB bandwidth as wide as possible to minimize

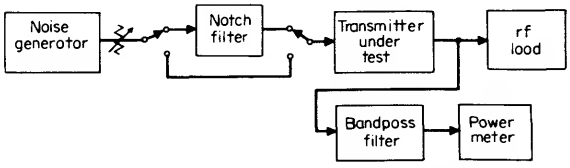


FIG 14-11 Essential block diagram of noise-loading test for single-sideband transmitters.

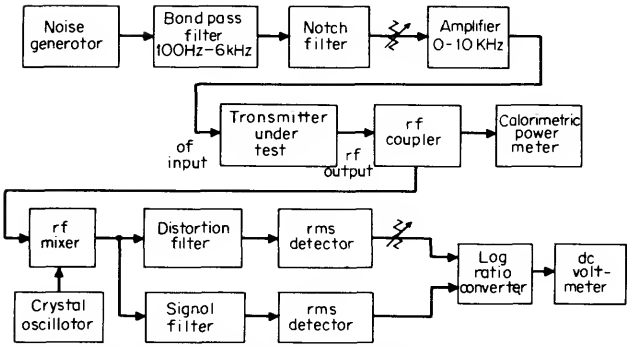


FIG 14-12 Block diagram of automatic noise-loading test set with continuous display of single-sideband noise power ratio.

the averaging time of the rms detector. A 3-dB bandwidth of 50 Hz is typical. The 60-dB bandwidth must be quite narrow (approximately 100 Hz) so that noise power outside the notched band does not limit the measurement dynamic range to less than 50 dB. A crystal filter, typically with a 100-kHz center frequency, is necessary for the distortion filter in order to meet these selectivity requirements. The signal filter is not strictly necessary since the loss of total power due to the notch can be calibrated out. When used, however, the signal filter permits readings to be taken of noise power ratio (single-sideband) which are relatively independent of output noise signal level. The signal filter is also a crystal bandpass filter whose center frequency is about 500 Hz removed from the distortion-filter center frequency. Its 3- and 60-dB bandwidths are typically 75 and 150 Hz. The rf coupler is any network which will couple a small sample of the transmitter output power to the following rf mixer. Since the signal and distortion filters are at or near 100 kHz, the local oscillator frequency is equal to that of the noise notch at the transmitter output ± 100 kHz. For example, if the single-sideband-transmitter suppressed-carrier frequency is 4,000 kHz and a single upper sideband is selected, then the noise notch is at 4,001.6 kHz and crystals at 3901.6 or 4101.6 kHz may be used. The use of a crystal oscillator for the frequency translation is preferable since the additive noise from the frequency translation will be the least with that type of oscillator. A frequency translation is required because the narrow filter bandwidths are not generally obtainable higher in the high frequency range.

The setup of Fig. 14-12 is initially adjusted by temporarily removing the notch filter and inputing flat spectral noise to the transmitter. The gain of the distortion filter channel is then adjusted with an attenuator such that the dc voltmeter reads full scale or at some other reference

reading. (The particular constants of the logarithmic ratio converter used must be known to calibrate the dc voltmeter.) The notch filter is reinserted and the noise power ratio (single-sideband) is read on the dc voltmeter. The calorimetric power meter indicates the transmitter noise power output directly, and the attenuator following the notch filter may then be used to vary the power level from zero to some point above rated noise power while taking readings of single-sideband noise power ratio.

When noise-loading distortion tests are made on a multichannel single-sideband transmitter, it is usual to input noise to all channels, putting the notch filter in the input to one of the two channels adjacent to the suppressed carrier. Separate noise generators must be used for each channel so that the output stages of the transmitter are amplifying a noise signal with Rayleigh amplitude distribution formed from the linear addition of separate sources. If a single noise source were used for all channels, the rf envelope of each sideband channel would contain peaks which occur simultaneously, that is, they would be correlated.

The characteristics of the channel noise source should be measured to affirm that the noise is gaussian within certain limits. The spectral output should be flat with frequency within ± 1 dB over the band of 100 to 6,000 Hz. The amplitude density distribution should follow a normal distribution within the limits of the table given here.

Voltage, σ	*Amplitude density† distribution*
0	0.0796 ± 0.005
±1	0.0484 ± 0.005
±2	0.0108 ± 0.003
±3	0.000898 ± 0.0002

† Amplitude density is measured in a window of 0.2σ, where σ is the standard deviation or rms value of the noise voltage.

A representative performance requirement for a single-sideband transmitter can be a 40-dB single-sideband noise power ratio measured at rated noise power.

Differential Time Delay. In certain types of communication systems, notably those handling a multitude of data signals, a parameter of great importance is the relative time delay between signals at different frequencies in the reception (or transmission) passband. Ideally, this relative delay should be zero so that at the output, the original waveform is reconstructed without distortion. This follows from the Fourier decomposition of a complex function of time; the original waveform is obtained from summation of the components when both the relative amplitude and phase are preserved.

This concept may be further illustrated with Fig. 14-13 in mind. Suppose a complex waveform is passed through a linear network from A to B. The sinusoidal components at point A must have a certain time relationship to produce the complex wave; for example, component A_1 passes through zero at time t_1 with positive slope, and A_2 passes through zero with positive slope at t_2. Now, if an observer at point B is to see the same complex waveform, the interval between positive-slope zero crossings at point B must be exactly the same as those at point A. If there is an absolute time delay T in the network, the component B_1 must cross zero at $T + t_1$ and the component B_2 must cross zero at $T + t_2$. Thus the value T must be independent of frequency.

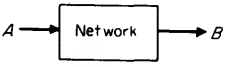


FIG 14-13 Representation of a network with time delay passing a complex wave from A to B.

Absolute time delay may be thought of as phase delay. If a sinusoid with a frequency of 1 Hz enters point A of Fig. 14-13 and suffers a time delay of 1 sec by passing through the network to point B, it may be seen that the equivalent phase retardation is 360° or 2π rad. If the phase retardation were at 2 Hz and the delay still 1 sec, the phase retardation would be 720° or 4π rad, and so forth. Therefore, the absolute time delay or phase delay is $T_{abs} = \beta/\omega$, where β is the phase shift in radians and ω is the angular frequency in radians per second ($2\pi f$). This relation is used in sketching the relationship between constant absolute time delay and the corresponding phase shift versus frequency in Fig. 14-14. Note that for the time delay to be constant versus frequency, the phase shift must be linear with frequency and have a slope equal to the delay. Additionally, each component frequency must be attenuated (or amplified) by an exactly equal amount if there is to be no distortion of the waveform after passing through the network.

Communication systems are, by nature, band limited; that is, only a certain range of frequencies is passed by the receiver or transmitter. For example, a single-channel single-sideband system operating at a sup-

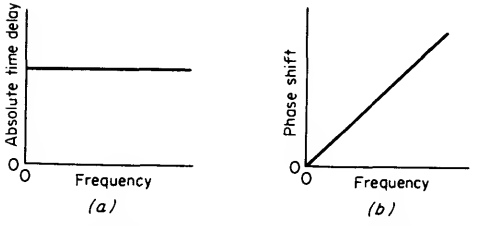


FIG 14-14 The relationship between time delay and phase shift of a wave.

pressed carrier frequency of 2 MHz (upper sideband used) may have a usable band extending from 300 to 3,000 Hz above the carrier. For distortionless transmission it is only necessary that the amplitude response be flat and the phase response linear with frequency over the band 2.0003 to 2.003 MHz. Furthermore, with respect to time delay, it is sufficient that the relative or differential time delay over this band of frequencies be zero. The absolute time delay can be any amount; as an extreme example, the absolute time delay could be several seconds for an interplanetary communication link. The capability of a single-sideband data receiver or transmitter to reproduce a complex waveform faithfully, then, may be characterized by measuring versus frequency both the relative amplitude response and the differential time delay.

In a band-limited system, the relative or differential time delay at a given frequency is the slope of the phase β versus frequency ω. Thus, the differential time delay T_d is

$$T_d = \frac{d\beta}{d\omega} \tag{14-9-1}$$

This quantity is also referred to in the literature as *group* or *envelope delay*, since a group of closely spaced frequencies having a certain rf envelope will be almost equally delayed by the amount T_d. It is evident that measurement of the quantity $\Delta\beta/\Delta\omega$, where $\Delta\beta$ is the change in phase for an increment in frequency $\Delta\omega$, will closely approximate T_d when $\Delta\omega$ is a suitably small increment of frequency.

The methods of measuring differential time delay were documented very early in the technical literature [4]. Broadly, two methods of measurement are possible: indirect calculation by measuring phase versus frequency, and the direct measurement of the delay of an envelope through the system. Commercially, the latter method is usually now employed and may be basically explained in the following way: Sketched in Fig. 14-15 is an example of the phase response of a system whose phase is assumed linear with frequency over some small band of frequencies. Suppose that an AM carrier ω_0 is passed through the system. If the modulating frequency is ω_m, then the sideband frequencies ω_1 and ω_2 are $\omega_0 - \omega_m$ and $\omega_0 + \omega_m$, respectively. Upon passing through the system, these sidebands are rotated with respect to the carrier by the angles $\beta_0 - \beta_1$ (lower sideband) and $\beta_2 - \beta_0$ (upper sideband). Since linear phase is assumed, these rotations

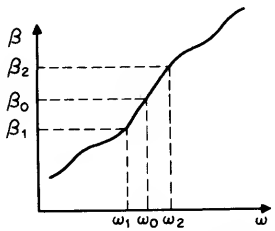


FIG 14-15 Example of phase behavior of a system having linear phase response in the region $\omega_1 < \omega < \omega_2$.

are equal and may be called $\Delta\beta$. The corresponding frequency change $\Delta\omega$ producing $\Delta\beta$ is ω_0. Therefore, the envelope delay at ω_0 is approximated by

$$T_d = \frac{\Delta\beta}{\omega_m}$$

where $\Delta\beta$ is measured in radians and ω_m in radians per second. Correspondingly,

$$T_d = \frac{\Delta\phi}{360 f_m}$$

where $\Delta\phi$ is in degrees and f_m is the modulation frequency in hertz. These equations say that the time delay may be measured by finding the input-to-output phase shift ($\Delta\beta$ or $\Delta\phi$) of the modulation *envelope* and dividing by the modulation frequency.

It is to be noted that a real system does not have $\beta_2 - \beta_0$ equal to $\beta_0 - \beta_1$. Therefore, in general, the output envelope does not exactly have the shape of the input envelope. A compromise is necessary in the choice of ω_m such that $\Delta\beta$ is accurately discernible (high ω_m) and the envelope distortion is small (low ω_m) so that like points (peak or valley) may be compared at the input and output.

Figure 14-16 shows a test setup that can be used to measure the differential time delay of a transmitter. The rf signal generator is modulated at a low audio frequency, 30 to 50 Hz, by an external oscillator. The output of the rf signal generator is applied to the transmitter under test at a point prior to any frequency-selective filters. In the case of a single-sideband transmitter, this point would be at the output of the balanced modulator sideband generator where the carrier frequency would typically be in the 100- to 455-kHz region. Both the input and output rf signals are demodulated by identical envelope detectors whose time constants should be short compared with the modulation frequency ω_m. The

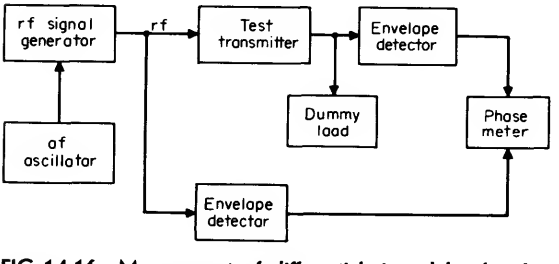


FIG 14-16 Measurement of differential time delay by the direct method of envelope delay.

envelope signals are then compared in phase as a function of the rf signal in the passband of the transmitter. The phasemeter can be an instrument designed for that purpose, or basically, an oscilloscope may be used by obtaining a Lissajous pattern. The accuracy and resolution requirements of the phase comparison are quite high; a 5.4° change in phase $\Delta\phi$ is equivalent to a 500-μsec change in delay T_d when the modulation frequency f_m is 30 Hz. The accuracy of the envelope detectors may be checked by interchanging them and repeating the measuring run. The average of the two runs is then due to the delay of the transmitter under test. It should be noted that this measurement really yields only the differential time delay because the angle indicated by the phasemeter may be the residue of one or more 360° rotations. Since differential time-delay distortion is the important parameter affecting complex signal transmission, this is not a serious drawback. The absolute time delay can be found by decreasing f_m until it is apparent that the total phase rotation through the transmitter is less than 360°. The results of a differential-time-delay measurement on a single-sideband transmitter are shown in Fig. 14-17, which is plotted to show the absolute time delay. The transmitter tested has delay equalization networks added to the sideband selection filter. The measurement was made between the audio input terminals of one channel and the 250-kHz output of the transmitter channel selection filter.

Test equipment designed expressly for measuring differential time delay and also the attenuation versus frequency of a system is available from manufacturers such as Wandel and Goltermann (West Germany) and the Hewlett-Packard Company. Such equipment uses the Nyquist-Brand

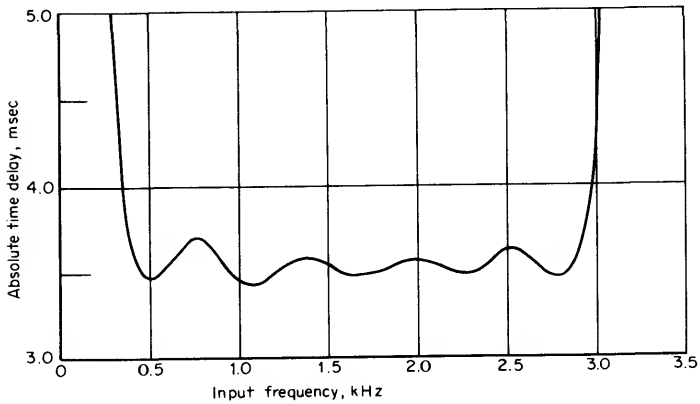


FIG 14-17 Time delay of a single-sideband transmitter.

measurement principles with modifications allowing swept frequency display of attenuation and delay. The Wandel and Goltermann LD-2 equipment uses a two-carrier differential measurement technique in which a reference rf envelope is first passed through the test specimen at approximately its band-center frequency. Then the reference signal is replaced by a slowly swept FM carrier. This changeover is effected, without disruption of the rf envelope, at a 4-Hz rate. At the receiving end, the envelope phase change is measured and converted to time delay. This method allows differential delay measurements on communication systems whose input and output terminals are not in the same place.

Residual AM Noise. When the modulation input to an AM or FM transmitter is removed, the variations in the rf envelope amplitude are called *residual AM noise.* The internal sources of such noise are power supply hum, audio wiring pickup of stray fields, and the thermal and excess noise of the active stages in the transmitter. Although hum is roughly sinusoidal, it is termed noise in the context that it is an undesired signal.

To perform the measurement, an AM detector responsive to a linear peak carrier is coupled to the output, or a sample of the output, of the transmitter. For an AM transmitter, the ratio between detected audio for 0 and 100 percent modulation is the residual noise level and is usually expressed in decibels. For FM transmitters, the same detector is used and, with no frequency modulation, the ratio between the peak value of the envelope and the average detected dc voltage is the residual AM noise level. In both cases, when no modulation input to the transmitter is present, the audio terminals should be terminated by a resistor equal to the rated audio-source resistance.

The measurement of residual noise in an AM transmitter is very simple when a harmonic distortion analyzer is used which incorporates an integral diode detector. Such instruments are often used in AM broadcast-station monitoring equipment to additionally measure audio distortion on the rf envelope, the distortion being brought into the instrument from a sampling loop mounted on the antenna.

Residual FM Noise. This measurement is similar to the AM noise test except that an FM detector is coupled to the output of the transmitter. The FM detector may be a deviation meter or an FM receiver whose quieting characteristic is greater than the FM noise level intended to be measured. For FM transmitters, the FM detector output ratio for full-rated system deviation and no audio input is the residual FM noise level. For AM equipment, the FM detector output, both with and without AM, is measured with the reference being the FM detector output with an FM input signal having some standard amount of deviation, such as 1,000 Hz.

Keying Waveshape. The transmission of on and off signals, such as international Morse code, requires that the output of the transmitter be turned on and off in accordance with the code. If the transition time from off to on or reverse is very abrupt, sideband energy is produced which extends several kilohertz about the carrier frequency and results in interference to services occupying adjacent frequencies. This follows from the general equation of an AM wave

$$e(t) = E[1 + m(t)] \cos \omega_c t \qquad (14\text{-}9\text{-}2)$$

where $e(t)$ = instantaneous output-voltage function of time
E = peak carrier amplitude
$m(t)$ = modulating function of time
ω_c = angular carrier frequency

When $m(t)$ is a square wave, an infinite set of sidebands appears about the carrier, displaced in frequency from the carrier by integral odd multiples of the square-wave frequency. On the other hand, if the rise and fall times are too slow, the code elements tend to merge at high sending rates, which makes recognition of manual or electronic codes difficult.

Measurement of keying waveshape may be made in either the frequency or time domains. In the frequency domain, direct measurement of the transmitter spectrum can be obtained by applying a sample of the transmitter output to a spectrum analyzer and using the signal analysis techniques of Chap. 5. To obtain a stationary spectrum, the keying input to the transmitter must be a constant frequency. Since the transmitter normally provides waveshaping, the keying input may therefore be a square wave of frequency f equal to the highest expected keying rate. For international Morse code, $f = 0.416$ words per minute, where f is in hertz and words per minute is the keying rate.

In the time domain, keying waveshape may be determined by the setup shown in Fig. 14-18. The square-wave generator frequency can be adjusted over the expected range of keying rate while the leading and trailing edges of the transmitter output are observed. Although the direct connection of a high-frequency oscilloscope is shown in Fig. 14-18, a

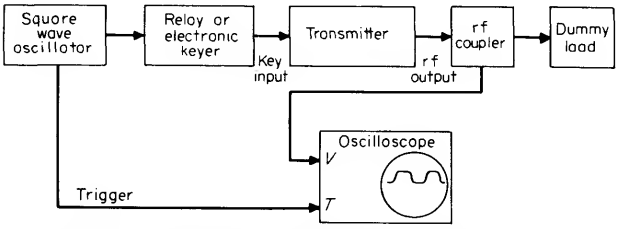


FIG 14-18 Keying waveshape time-domain measurement.

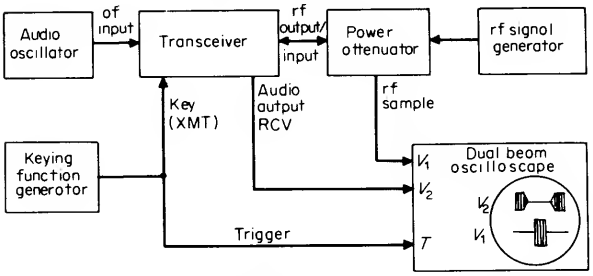


FIG 14-19 Turnaround-time measurement with a dual-beam oscilloscope.

diode detector may be interposed if the frequency capability of the oscilloscope is insufficient. The load time constant of such a detector, however, must be short compared with the transmitter rise and fall time. For there to be little energy in the higher-order sidebands, the time display should show smoothly rounded leading and trailing edges. The transmitter should rise to full power, however, at the highest keying rate. A noteworthy byproduct of the time display is observation of amplitude in the interval between leading and trailing edges. Droop in this interval is caused by regulation of the output amplifier power supply. At certain keying rates, a damped oscillation may be noted, caused by resonance of the power supply filter.

Turnaround Time. This measurement is included as an example of one of the many specialized performance tests for particular classes of radio communication equipment and has to do with the amount of time required for a transceiver (combination transmitter and receiver) to go from "transmit" to "receive." Rapid turnaround time is a requirement for certain two-way data systems (usually high-frequency single-sideband) where transmission of a bit stream is quickly followed by reception of another transmission.

A method of measuring the transfer time between reception and transmission and vice versa by using a dual-beam oscilloscope is sketched in Fig. 14-19. In this case, the transceiver is assumed to be a single-sideband type such that a single rf output is generated by the transmitter section when a single audio frequency is applied to the audio input terminals. The keying-function generator can be simply a relay actuated repetitively by a low-frequency square-wave oscillator. The horizontal sweep of the oscilloscope is triggered by either the leading or trailing edge of the transmit key signal; leading-edge triggering is assumed for the sample waveforms depicted. The receiver audio output is applied to one vertical axis input, and a sample of the rf output is connected

to the other vertical input. A power attenuator is shown between the rf output-input of the transceiver and an rf signal generator; this attenuator must dissipate the output power of the transmitter and have enough loss to protect the signal generator from burnout due to the transmitted signal. Alternatively, an rf coupler can be employed to inject the rf output of the signal generator into the transceiver with the use of a dummy load to dissipate the transmit power.

The test is performed by keying the equipment at a rate slow enough that complete receive-transmit changeover occurs and then adjusting the horizontal-sweep rate so that both changeover intervals may be seen on the oscilloscope. With the sweep-rate calibration, the transfer times may be directly read off the display. Transfer times of 50 to 100 msec are typical for single-sideband data equipment. The rf-signal-generator equivalent level at the transceiver antenna terminals may be adjusted to some level such as 100 μV. This may be accomplished by knowing the total attenuation between the rf signal generator and the transceiver and using the calibrated rf output attenuator of the rf generator. Alternatively, the desired rf level may be temporarily applied directly to the transceiver, and the internal automatic-gain-control voltage noted. With the rf signal generator again connected as in Fig. 14-19, the generator output attenuation can then be decreased until the same automatic-gain-control voltage is developed.

Communication System Interference. In installations with multiple transceivers and antennas operating within the same frequency band (but not on the same frequency), severe interference can arise when the isolation between antennas becomes low because of physical constraints on antenna placement. This situation is common in very high frequency voice-communication installations in aircraft, where the isolation between antennas can range from 20 to 50 dB depending on the airplane size. The interference results from receiver overload, receiver spurious responses, and transmitter spurious output. The subjective effects are that a desired reception is garbled or even completely blocked by transmission from the nearby transceiver, and that the squelch of a transceiver in the receive condition can be falsely opened (as if a signal were present) by transmission from the other transceiver.

A test method is needed that will simulate, in the laboratory, the field conditions under which two transceivers are expected to operate, including the antenna isolation and the effects of transmitter and receiver spurious emissions and responses. Figure 14-20 shows one method of measurement that allows quantitative measurement of interference effects. Two transceivers are used; they are typically identical types of equipment in a dual installation. Transceiver A is operated in the transmit condition, and transceiver B in the receive condition. For explanatory pur-

poses, the transceivers will be assumed to be AM voice equipment. Transceiver A is modulated to its full capability (80 to 90 percent) at 400 Hz from an audio-frequency oscillator. The rf output is applied to a power attenuator capable of dissipating the transmitter power and

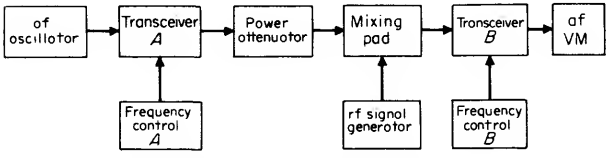


FIG 14-20 Communication system interference test.

having an insertion loss equal to the actual antenna isolation less 6 dB. An additional 6-dB loss is incurred in the mixing pad (which is identical with that discussed in connection with Fig. 14-7), which brings the total attenuation between transceivers equal to that between antennas in the installation. The audio output of transceiver B is monitored with an audio-frequency voltmeter. (Headphones or a loudspeaker at this point are a useful accessory to identify the character of the interference.) An rf signal generator modulated 30 percent at 1,000 Hz is also applied to the mixing pad to simulate a received signal.

Two tests can be made with the setup in Fig. 14-20. The first test is a squelch-opening test and is performed with no signal from the rf signal generator and with the squelch control of transceiver B adjusted so that a signal of typically 5 μV is required to unmute the receiver section. Frequency control B is set for some desired channel within the operating band of frequencies. With transceiver A transmitting, frequency control A is exercised through all channels of the band and those channels noted which cause the squelch of transceiver B to open as indicated by the audio-frequency voltmeter or other audio output indicator. The test may be repeated for as many receiver channel settings of frequency control B as are required to characterize the system. The causes of squelch openings are transmitter envelope distortion resulting in components lying on the receiver channel frequency, inadequate receiver selectivity, transmitter spurious outputs, and receiver spurious responses. The last are usually predominant and arise chiefly from the first mixer in a superheterodyne receiver. They may be identified by using the general expression for a frequency-converting mixer

$$Hf_H + Lf_L = f_0 \tag{14-9-3}$$

where H and L are integers of either sign associated with f_H, the higher-frequency input, and f_L, the lower-frequency input, and f_0 is the mixer

output frequency or the intermediate frequency in a superheterodyne receiver. For example, suppose that the desired signal frequency is 120 MHz (f_L) and that the local oscillator is at 140 MHz (f_H). The desired mixing product is obtained when $H = +1$ and $L = -1$, which yields an f_0 of 20 MHz. Now suppose that an undesired signal is present at 130 MHz. By trial and error or the use of either spurious-prediction charts [5] or a computer program [6], it can be found that when $H = +2$ and $L = -2$, then f_0 again equals 20 MHz. The sum of the absolute values of H and L is termed the *order of the spurious response*, in this case, 4. Spurious responses as high as thirtieth order have been observed to produce squelch openings in typical very high frequency transceivers when the antenna isolation is as low as 20 dB. The gain of the mixer relative to the desired mixing process decreases markedly as the order increases. Modern mixers with hot carrier diodes or field effect transistors typically exhibit spurious ratios of -60 to -80 dB for orders above the tenth.

The second test performed with the equipment in Fig: 14-20 is closely akin to a standard crossmodulation or blocking test. Transceiver B is channeled to a desired frequency in its range, and the rf signal generator is used to apply perhaps a 20-μV signal. For very high frequency aircraft transceivers, this approximates the minimum signal received while in flight; the test can be repeated for other desired signal levels if required. With transceiver A transmitting and modulated, frequency control A is used to channel the transceiver over its frequency range. At each channel, a measurement of signal plus noise to noise is made on transceiver B by removing the modulation from the rf signal generator and noting the drop in audio output on the audio-frequency voltmeter. This test can be very time consuming if a large number of channels are involved; it is often sufficient to cover the range of transceiver A in 1-MHz steps. The results of the test combine the interference effects of discrete transmitter and receiver spurious frequencies, crossmodulation and blocking of the receiver, and wideband noise output of the transmitter. The data can be plotted as the ratio of signal plus noise to noise in the frequency spacing between transceivers A and B. A practical minimum ratio of signal plus noise to noise is 10 dB, at which point the interference is annoying and speech intelligibility is beginning to suffer. A 20-dB ratio of signal plus noise to noise subjectively is nearly unnoticeable, particularly in an acoustically noisy environment.

14-10 Radio Equipment Specifications

The successful design, construction, and measurement of radio receiving and transmitting equipment require detailed knowledge of the per-

formance specifications applicable to the particular class of equipment. There are a great number of published specifications which apply to commercial and military radio equipment; those most frequently encountered are listed in this section, together with instructions for obtaining copies. These specifications generally cover the minimum performance requirements for an equipment. Certain documents, notably those of the Federal Communications Commission (FCC) and the Radio Technical Commission for Aeronautics (RTCA), define the legal minimum specifications that are required for sale and operation of radio equipment in the United States. The actual performance of radio receivers or transmitters usually must exceed these minimum values to be competitive in the electronics market.

Commercial Specifications. The following groups of specifications are listed, along with the organization from which copies may be obtained. The list is confined to those specifications concerned with radio receiving or transmitting equipment used for communication or navigation purposes.

Electronics Industries Association
2001 Eye Street, N.W.
Washington, D.C. 20006

Electronics Industries Association (EIA) specifications are mainly written to define the performance of two-way FM radio equipment for services such as radio for business, police, fire, taxicabs, forestry, and personal paging. The actual performance of such equipment is usually much better than the minimum standards specified.

Publication	*Title*
RS-152-A	Land Mobile Communications, FM or PM Transmitters (25 to 470 mc[1])
RS-204	Minimum Standards for Land-mobile Communication FM or PM Receivers
RS-237	Minimum Standard for Land-mobile Communication Systems Using FM or PM in the 25–470 mc Frequency Spectrum
RS-240	Electrical Performance Standards for Television Broadcast Transmitters
RS-250A	Electrical Performance Standards for Television Relay Facilities
RS-316	Minimum Standard for Portable/Personal Land Mobile Communications FM or PM Equipment 25–470 mc
TR-107	Electrical Performance Standards for FM Broadcast Transmitters (88 mc–108 mc)
TR-120	Minimum Standards for Land-mobile Selective Signaling Equipment

[1] mc = MHz.

Aeronautical Radio, Incorporated
2551 Riva Road
Annapolis, Maryland 21401

Aeronautical Radio, Incorporated (ARINC) is an association supported by the airlines of the United States. Specifications of ARINC therefore cover the equipment needed for radio communication and navigation in commercial aircraft. These specifications define the actual operating requirements rather than the legal minimum performance; in certain areas, notably crossmodulation and intermodulation performance of receivers, the requirements may represent goals which are difficult to obtain in production equipment.

Publication, characteristic No.	*Title*
533A	Airborne HF[1] SSB[2]/AM System
546	Airborne VHF[3] Communications Transceiver System
547	Airborne VHF Navigation Receiver
551	Airborne Glide Slope Receiver—Mark 2
552	Radio Altimeter
566	Airborne VHF Communications Transceiver and Mark-1 VHF Satcom[4] System
568	Airborne Distance Measuring Equipment
570	Mark-3 Airborne ADF[5] System
572	Mark-2 Air Traffic Control Transponder

[1] HF = high frequency.
[2] SSB = single sideband.
[3] VHF = very high frequency.
[4] Satcom = satellite communications.
[5] ADF = automatic direction finder.

Underwriters Laboratories, Incorporated
207 East Ohio Street
Chicago, Illinois 60611

1285 Walt Whitman Boulevard
Melville, L.I., New York 11746

1655 Scott Boulevard
Santa Clara, California 95050

The following electrical safety specification is included to give a guide for the proper construction of radio equipment for consumer service. It is available without cost from any of the above addresses.
Standards for Safety, Radio, and Television Receiving Appliances, 492.

Federal and Military Specifications. Rules, regulations, and specifications written by agencies of the United States government are obtainable either from the U.S. Government Printing Office, Washington, D.C. 20402 or the sponsoring agency; those listed below apply to commercial

and military uses of radio. Each branch of the military also originates specifications for their specific requirements; these may be obtained (some with restrictions) from the U.S. Department of Defense, Washington, D.C. 20360.

Federal Communications Commission
(Documents available by volume from U.S. Government Printing Office
Washington, D.C. 20402)

The following volumes and parts thereof define the FCC regulations which must be adhered to in the performance and use of radio equipment in the United States. These regulations chiefly cover any equipment which emits rf energy; receiver performance is only specified with regard to sensitivity (noise figure) and frequency range for television receivers and antenna radiation limits.

Volume	*Part*	*Title*
2	2	Frequency Allocations and Radio Treaty Matters; General Rules
2	5	Experimental Radio Services
2	15	Radio Frequency Devices
3	73	Radio Broadcast Services
3	74	Experimental, Auxiliary, and Special Broadcast and Other Program Distributional Services
4	81	Stations on Land in the Maritime Services
4	83	Stations on Shipboard in the Maritime Services
4	85	Public Fixed Stations and Stations of the Maritime Services in Alaska
5	87	Aviation Services
5	89	Public Safety Radio Services
5	91	Industrial Radio Services
5	93	Land Transportation Radio Services
6	95	Citizens Radio Service
6	97	Amateur Radio Service
6	99	Disaster Communications Service
7	21	Domestic Public Radio Services
7	23	International Fixed Public Radiocommunication Services
7	25	Satellite Communications

Radio Technical Commission of Aeronautics
2000 K Street, N.W.
Washington, D.C. 20006

The RTCA is a United States government agency concerned in part with establishing the legal minimum performance of radio equipment used by aviation in the United States. These standards must be met in order to obtain type certification of an equipment for use in commercial aircraft. Copies may be obtained directly from the above address.

Publication	*Title*
DO-48A	Minimum Performance Standards—Airborne Radio Communication Transmitting Equipment Operating within the Radio-frequency Range of 1.5–30 Megacycles
DO-49A	Minimum Performance Standards—Airborne Radio Communication Receiving Equipment Operating within the Radio-frequency Range of 1.5–30 Megacycles
DO-57A	Minimum Performance Standards—Airborne Radio Marker Receiving Equipment Operating on 75 Megacycles
DO-86	Characteristics of Aeronautical Single-sideband Systems
DO-92	Minimum Performance Standards—Airborne Loran A Receiving Equipment Operating within the Radio-frequency Range of 1,800–2,000 Kilocycles
DO-93	Minimum Performance Standards—Airborne Selective Calling Equipment
DO-94	Minimum Performance Standards—Portable aircraft Emergency Communications Equipment Operating within the Radio-frequency Range of 118–250 Megacycles
DO-95	Minimum Performance Standards—Portable aircraft Emergency Communications Equipment Operating within the Radio-frequency Range of 450–8,500 Kilocycles
DO-109	Minimum Performance Standards—Airborne Radio Communications Receiving Equipment Operating within the Radio-frequency Range of 117.975–136.000 Megacycles
DO-110	Minimum Performance Standards—Airborne Radio Communications Transmitting Equipment Operating within the Radio-frequency Range of 117.975–136.000 Megacycles
DO-111	Minimum Performance Standards—Airborne Receiving and Direction Finding Equipment Operating in the Radio-frequency Range of 200–400 Kilocycles
DO-112	Minimum Performance Standards—Airborne ATC[1] Transponder Equipment
DO-114	Minimum Performance Standards—Airborne VOR[2] Receiving Equipment Operating within the Radio-frequency Range of 108–118 Megacycles
DO-123	Minimum Performance Standards—Airborne Low-range Radar Altimeters
DO-124	Minimum Performance Standards—Airborne Distance Measuring Equipment (DME) Operating within the Radio-frequency Range of 960–1,215 Megacycles
DO-132	Minimum Performance Standards—Airborne ILS[3] Glide Slope Receiving Equipment

[1] ATC = air traffic control.
[2] VOR = very high frequency omnirange.
[3] ILS = instrument landing system.

Military Specifications. The three military services write specifications relating to every area of military procurement. Those specifications prefixed with *MIL* are available from the Department of Defense, Washington, D.C. 20360.

The first two specifications are noteworthy in that they describe the measurement of and requirements for EMI. The third specification defines communication system requirements and closely parallels the information in DCA-CIR-175-2A (see the next following reference) on high-frequency single-sideband transmitters and receivers.

Publication	*Title*
MIL-STD-461	Electromagnetic Interference Characteristics—Requirements for Equipment
MIL-STD-462	Electromagnetic Interference Characteristics, Measurement of
MIL-STD-188C	Military Communication System Technical Standards

Defense Communications Agency. The following specification covers military communication systems and equipment from very low to very high frequency. It is particularly excellent in the area of high-frequency single-sideband transmitters and receivers, giving design requirements that are near the most advanced developments. It is obtainable through the U.S. Government Printing Office.

Publication	*Title*
DCA-CIR-175-2A	DCS[1] Engineering-installation Standards Manual

[1] DCS = defense communication systems.

CITED REFERENCES

1. Gannaway, R.: Signal-to-noise Ratio in Receivers Using Linear or Square-law Envelope Detectors, *Proc. IEEE Letters*, October, 1965.
2. Fubini, E. G., and D. C. Johnson: Signal-to-noise Ratio in AM Receivers, *Proc. IRE*, vol 36, pp. 1461–1466, December, 1948.
3. Van Valkenburg, M. E.: "Introduction to Modern Network Synthesis," John Wiley & Sons, Inc., New York, 1964.
4. Nyquist, H., and S. Brand: Measurement of Phase Distortion, *Bell System Tech. J.*, vol. 9, 1930.
5. Brown, T. T.: Mixer Harmonic Chart, *Electronics*, April, 1951; or Olsen, W. R., and R. V. Salcedo: Mixer Frequency Charts, *Frequency*, March–April, 1966.
6. Myers, R. T., and T. A. McKee: Receiver Spurious Responses—Computer Improves Receiver Design, *IEEE Trans. Vehicular Commun.*, March, 1966.

CHAPTER FIFTEEN

MICROWAVE SIGNAL SOURCES

From notes by

Stephen F. Adam

John J. Dupre

Douglas Gray

Steve Hamilton

Wallace Rasmussen

Hewlett-Packard Company
Palo Alto, California

and

L. Besser

Fairchild Electronics Corporation
Mountain View, California

The importance of having adequate sources of test signals for measurement procedures must be clear in the reader's mind by now. The development of excellent audio sources and rf sources for laboratory use has kept pace with user requirements rather easily through the years, but

the development of microwave sources has been more difficult. As the operating frequencies increased, lumped circuit components no longer appeared to be lumped, and conventional amplifying devices no longer followed conventional behavior. Distributed-impedance parameters limited or even prevented performance.

For years the most effective strategy in the face of these difficulties has been to make both the circuit components and the active devices distributed by their very nature. Reactances became sections of transmission lines of appropriate characteristic impedances, lengths, and kinds of termination. Amplifying devices began to depend upon the finite electron velocities within them, and so klystrons and traveling-wave tubes were born. These tubes performed because of finite transit times and finite velocities of charges, rather than in spite of these phenomena.

The theory of operation of these high-frequency electron tubes—klystrons, magnetrons, traveling-wave tubes, and backward-wave tubes—is beyond the scope of this book. Where the tubes appear in signal sources, it will be assumed that the reader has some knowledge of them. They will be described only to the extent necessary to clarify the salient points of the signal generators.

In this chapter, the term *signal source* is used in a general way. It can mean *an instrument that is a source of signals*, but it is usually applied to the basic device or circuit that converts dc power to ac power. In the latter sense it is synonymous with *oscillator*. We reserve a special meaning for the term *signal generator:* an instrument that produces test signals of very accurately known character. A standard-signal generator is calibrated and stable.

15-1 Microwave Transistor Oscillators

The special microwave tubes mentioned above have been extremely valuable in laboratory signal sources, but they have the following disadvantages: limited operating life of a few thousand hours or less, large physical size, and operating voltages in the kilovolt range. Some must be operated in a constant and accurately specified magnetic field. Furthermore, accurate AM and FM are difficult to achieve in some oscillator tubes.

Therefore, research has continued through the years on solid-state devices to replace the tubes. In particular, microwave transistors and their oscillator-operating circuits have extremely long life, can be made physically small through the use of hybrid microcircuit techniques, and require typical operating voltages of less than 50 V.

Transistor Characterization. The most useful figure of merit for an oscillator transistor is its maximum frequency of oscillation, f_{max}. Mason [1]

and many others have shown that this frequency occurs where the unilateral power gain of the device is equal to unity. In terms of the scattering matrix S of the device, the unilateral gain [2] is given by

$$U(S) = \frac{|s_{12} - s_{21}|^2}{\det (1 - S\dot{S})} \qquad (15\text{-}1\text{-}1)$$

where S is the complex conjugate of S. At microwave frequencies the scattering matrix is conveniently measured over wide frequency ranges, and U may be calculated to determine f_{max}.

Transistor equivalent circuits lose their usefulness in the microwave region because they are difficult to obtain and become very complex for accurate modeling. For oscillator design it is advantageous to use the measured scattering parameters directly and to avoid the intermediate step of determining an equivalent circuit. If another parameter set is more convenient for a particular design (for example, y or z parameters), computer conversion from scattering parameters to the new parameter set is straightforward.

Oscillator Starting Conditions. It is convenient to think of two basic oscillator types, the feedback oscillator and the negative-resistance oscillator, even though feedback oscillators exhibit negative resistance across certain terminals. The feedback oscillator consists of an amplifier and a frequency-selective feedback network connected, for instance, as in Fig. 15-1. A z connection (or series connection) of the amplifier and feedback networks is shown and is readily analyzed by using the z matrix of each network. The condition for oscillation is that the determinant of the overall circuit matrix be equal to zero [3], or

$$(z_{a11} + z_{f11})(z_{a22} + z_{f22}) - (z_{a21} + z_{f21})(z_{a12} + z_{f12}) = 0 \qquad (15\text{-}1\text{-}2)$$

where z_{aij} and z_{fij} are the z parameters of the amplifier and feedback networks respectively. An example of this type of oscillator is the Colpitts circuit shown in Fig. 15-2.

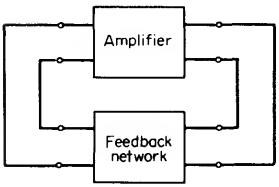


FIG 15-1 The z-connected feedback oscillator.

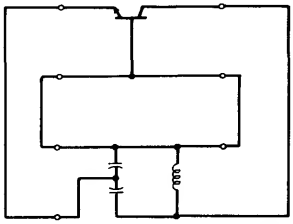


FIG 15-2 The Colpitts oscillator as an example of the z-connected circuit.

A negative-resistance oscillator consists of an active circuit exhibiting negative resistance at one port and a resonant network connected as in Fig. 15-3*a*. By letting Γ_R and Γ_A represent the reflection coefficients at the ports of the two networks, oscillatory growth will occur if at some frequency,

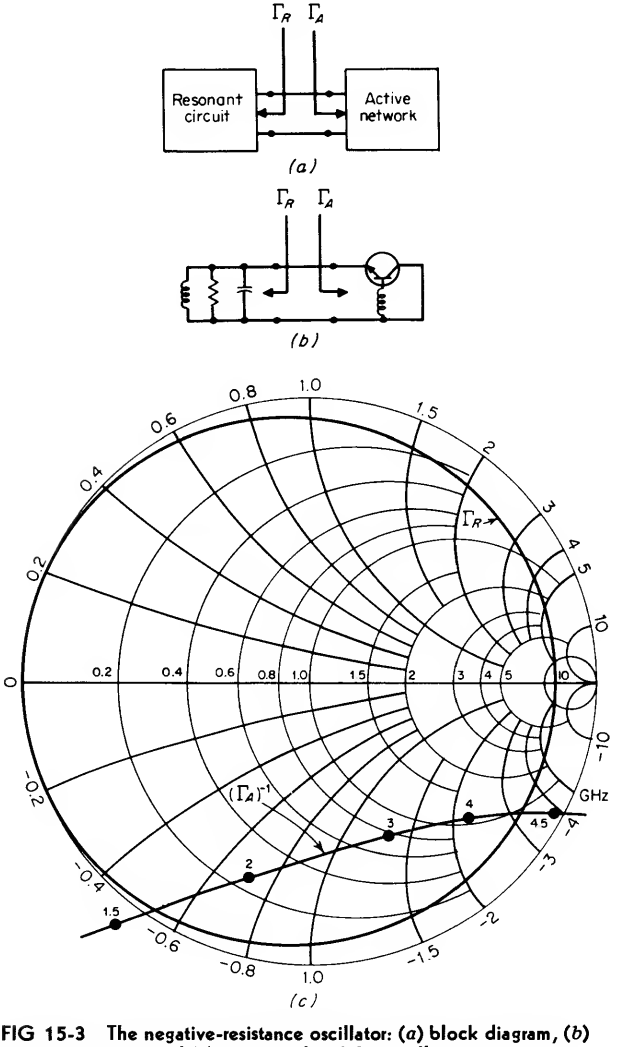


FIG 15-3 The negative-resistance oscillator: (*a*) block diagram, (*b*) a specific circuit, and (*c*) an example of the oscillation criterion.

$$|\Gamma_R| > \left|\frac{1}{\Gamma_A}\right| \quad \text{and} \quad \arg \Gamma_R = -\arg \Gamma_A \tag{15-1-3}$$

This implies that the Smith chart trace of Γ_R versus frequency must encircle $(\Gamma_A)^{-1}$ plotted on the same chart at the frequency of interest. As an example of this, consider the circuit in Fig. 15-3*b*. A negative resistance is provided at the emitter terminal of a common collector circuit with base inductance. The reciprocal of the reflection coefficient, $(\Gamma_A)^{-1}$, is plotted in Fig. 15-3*c* along with Γ_R, the reflection coefficient of the parallel resonant circuit. An oscillation range of approximately 1.7 to 4.3 GHz is apparent.

Oscillator Tuning. Oscillators for microwave sweep generators and spectrum analyzers are generally required to be electrically tunable over at least octave bandwidths. This may be accomplished with variable capacitance (varactor diodes) or yttrium-iron-garnet (yig) spheres.

Varactor diodes [4] have a capacitance approximately proportional to the square root of the applied reverse voltage. Thus, when they are used as part of a resonant network, the resonant frequency is proportional to the one-fourth power of the control voltage. Piecewise linear networks are required to linearize the voltage-versus-frequency relationship to better than within 1 percent. The ratio of maximum to minimum capacitance of a varactor is limited by the breakdown voltage and the parasitic capacitances, which makes the tuning of bandwidths exceeding one octave difficult to obtain.

In contrast, yig tuning is a very wideband tuning method. Filters with yig resonators and tuning from 1 to 12 GHz have been built. Oscillators that are yig tuned are usually limited in tuning range by the active circuit rather than by the resonator.

The yig resonance phenomenon takes place when a highly polished sphere of this ferrite material is placed in an rf structure under the influence of a dc magnetic field. A high-Q resonance occurs at a frequency proportional to the dc magnetic field. If the tuning field is supplied by an electromagnet, the resulting characteristic of magnet current versus resonant frequency is extremely linear (easily better than to within 0.1 percent over an octave). Linearizing networks are unnecessary. A typical coupling structure is a metallic loop surrounding the sphere as in Fig. 15-4*a*. The equivalent circuit looking into the loop terminals is that of a parallel resonant *RLC* circuit plus a series inductance (Fig. 15-4*b*). The values shown are for a gallium-doped yig sphere with a

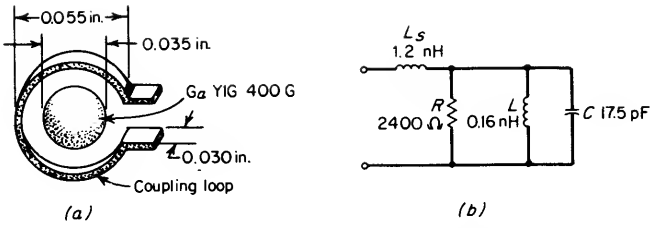


FIG 15-4 A typical yig resonator and its equivalent circuit.

saturation magnetization of 400 G and the coupling loop dimensions as shown.

Although yig tuning is inherently linear, broadband, and of high Q, the volume and power required by the tuning electromagnet are a disadvantage.

Example A 1.8- to 4.2-GHz yig-tuned Oscillator

Figure 15-5 shows the schematic diagram of a 1.8- to 4.2-GHz yig-tuned oscillator designed for instrument use. The oscillator is a negative-resistance type with its output amplified by a two-stage broadband amplifier.

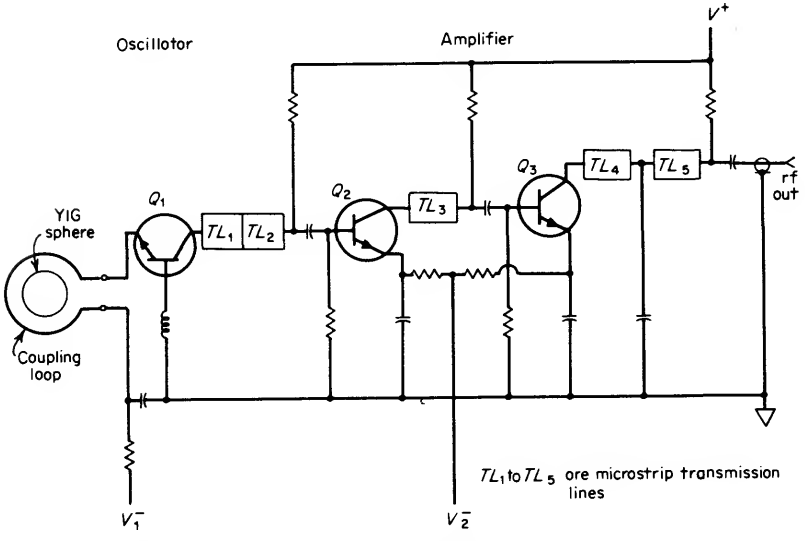
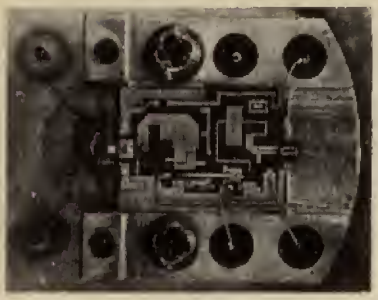


FIG 15-5 A 1.8- to 4.2-GHz oscillator and amplifier.



(a)

(b)

FIG 15-6 (*a*) The thin-film circuit in its holder. (*b*) The complete 1.8- to 4.2-GHz oscillator.

The amplifier buffers the oscillator from external load variations and boosts the output power to 20 mW across the frequency range. The oscillator and amplifier circuits are fabricated on a 0.5 × 0.750 in. sapphire substrate on which the passive elements have been deposited in thin-film form. Three transistor chips and the yig-sphere coupling loop are welded to the substrate. Figure 15-6 shows the thin-film circuit and the complete oscillator in its hermetically sealed package.

15-2 *Solid-state Microwave Amplifiers*

Since the 1940s until recently, traveling-wave tubes [5] have been the main active devices used in microwave amplifiers. For narrow-band applications requiring low noise, traveling-wave masers [6] and parametric amplifiers [7] were developed in the 1950s. Although these electron-tube devices are still used, solid-state devices have been rapidly replacing them, and this section only treats solid-state designs.

We specifically treat tunnel-diode [8] amplifiers and microwave transistor amplifiers and show how scattering parameters (*S* parameters) and

computer techniques are useful in amplifier design [9 to 12]. References 13 and 14 are especially recommended for those wishing to study the use of S parameters intensively in *design* work. The recent development of hybrid microcircuit technology has introduced a new era for the solid-state amplifiers. Volumes, weights, and circuit parasitics have been drastically reduced to yield higher performance and reliability. However, the benefits of the new technology can only be fully utilized when semiautomatic design techniques are used. The present availability of accurate network analyzers and digital-computer facilities has suddenly outdated the engineer who still wants to rely on "cut-and-try" techniques when designing microwave amplifiers.

Scattering parameters have been widely accepted to be one of the most powerful design and measurement tools. Since phase information is now also available, the designer is able to predict the behavior of his circuit very accurately, and by the help of computer optimization techniques, the ultimate performance can be approached. These design techniques are briefly illustrated below.

A brief comparison of the various types of microwave amplifiers is shown in Table 15-1.

Tunnel-diode Amplifier Design. When properly biased, the tunnel diode displays negative resistance that corresponds to a reflection coefficient of magnitude greater than 1, that is, the reflected wave is greater than the incident wave. If a directional-sensitive element such as a circulator is used, the incident and reflected waves can be separated and the tunnel diode can be used for amplification, and the available gain is directly proportional to the magnitude of the reflection coefficient. Theoretically, the tunnel diode would be an ideal device for broadband application, but the lack of broadband circulators typically limits the useful range of these amplifiers to an octave of bandwidth.

By basic definition, the reflection coefficient Γ of a one-port device as

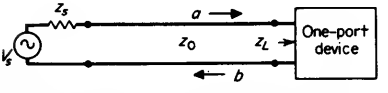


FIG 15-7 One-port device driven by a signal source.

shown in Fig. 15-7 is $\Gamma = b/a$. The voltage source is connected to the one-port through a transmission line to make $z_S = z_L = z_0$. Here a is a normalized voltage wave incident to the one-port network, and b is a normalized voltage wave reflected from the one-port network. When a

TABLE 15-1 Comparison of Microwave Amplifiers

Type of microwave amplifier	Advantages	Disadvantages
Traveling-wave tube.....	Broad bandwidth High-overload capability Usable to 100 GHz Wide dynamic range High gain and output power	High power consumption Limited life Warm-up required Large size
Traveling-wave maser....	Ultralow noise High gain stability Low intermodulation distortion	Large size High price Pump required with intrinsic frequency Intrinsic operating temperature
Parametric.....	Wide dynamic range Ultralow noise	Pump required High price
Tunnel diode.....	High gain per stage Small volume and weight Low noise Octave bandwidth available	Low output power Burnout protection required Difficult to stabilize Active device cannot be easily replaced
Transistor.....	Small volume and weight Broad bandwidth Low price Flat gain response High output power Wide dynamic range High reliability	Low gain per stage Currently limited to $f < 8$ GHz Burnout protection required
Monolithic integrated circuits	Ultimate reliability Smallest size Lowest cost	Currently limited to $f < 200$ MHz High noise figure

tunnel diode is biased in the negative resistance region, its reflection coefficient is considerably greater than unity. If a three-port circulator is connected to the tunnel diode as shown in Fig. 15-8, then a signal entering at port 1 will be amplified and will leave the circulator at port 3. A vector equation relating the normalized voltage waves and the s parameters of the circulator is

$$\begin{vmatrix} b_1 \\ b_2 \\ b_3 \end{vmatrix} = \begin{vmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{vmatrix} \begin{vmatrix} a_1 \\ a_2 \\ a_3 \end{vmatrix} \tag{15-2-1}$$

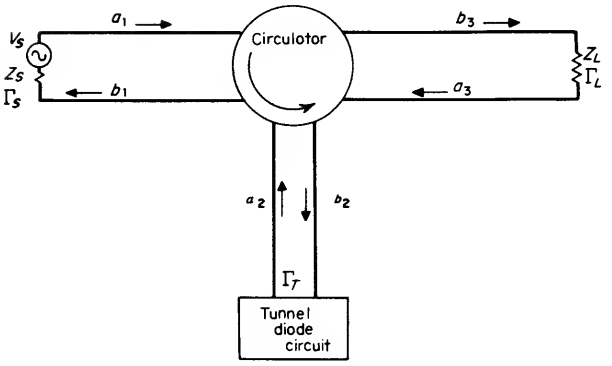


FIG 15-8 Simplified diagram of a tunnel-diode amplifier.

But, by definition, $a_2 = \Gamma_T b_2$ and $a_3 = \Gamma_L b_3$, or

$$\begin{vmatrix} b_1 \\ b_2 \\ b_3 \end{vmatrix} = |S| \begin{vmatrix} a_1 \\ \Gamma_T b_2 \\ \Gamma_L b_3 \end{vmatrix} \tag{15-2-2}$$

The lower-case letters will be used here to denote types of parameters and specific values of parameters. Capitals will denote whole matrices or networks.

The flow diagram of the above circuit is shown in Fig. 15-9. Note that the diagram is applicable to the general design when the source and load impedances are not necessarily equal to the characteristic impedance of the transmission lines.

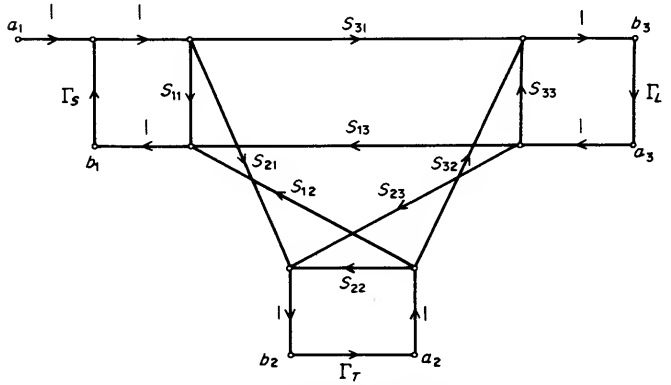


FIG 15-9 Signal flow diagram of a tunnel-diode amplifier.

Either by using the above vector relation or by applying a standard formula [15] to the flow diagram, the circuit can easily be analyzed. For example, the forward transducer gain of the amplifier is

$$G = \frac{b_3}{a_1} = \frac{s_{21}s_{32}\Gamma_T + s_{31}(1 - s_{22}\Gamma_T)}{1 - (s_{11}\Gamma_s + s_{22}\Gamma_D + s_{33}\Gamma_L + \Gamma_s\Gamma_T\Delta_1 + \Gamma_s\Gamma_L\Delta_2 + \Gamma_T\Gamma_L\Delta_3 + \Gamma_S\Gamma_T\Gamma_L\Delta_4)} \quad (15\text{-}2\text{-}3)$$

where $\Delta_1 = s_{21}s_{12} - s_{11}s_{22}$

$\Delta_2 = s_{31}s_{13} - s_{11}s_{33}$

$\Delta_3 = s_{32}s_{23} - s_{22}s_{33}$

$\Delta_4 = s_{11}s_{22}s_{33} + s_{31}s_{23}s_{12} + s_{13}s_{21}s_{32} - s_{11}s_{23}s_{32} - s_{22}s_{31}s_{13} - s_{33}s_{21}s_{12}$

In the above expressions, Γ_T applies to the tunnel diode including bias and stabilizing elements. The tunnel diode displays negative resistance from dc up to its cutoff frequency, and unless it works into a favorable reflection coefficient, the circuit may oscillate. Therefore, outside the useful range of the circulator, stabilizing elements should be used. These elements should not affect the in-band operation. A typical tunnel-diode amplifier circuit is shown in Fig. 15-10.

Transistor Amplifiers. The transistor amplifier is the most suitable kind for true broadband application. With the proper combination of feedback and impedance matching techniques, bandwidth can be extended to several octaves in the microwave region. Since the presently available active device will provide limited gain per stage, it is usually necessary to cascade several stages. A typical microwave-amplifier design is described below. The design is for a broadband, two-stage, 10-dB module used in a multistage 0.1- to 2.0-GHz amplifier [13].

To achieve flat gain through such a wide frequency range, the designer must use feedback at the lower frequencies and carefully match impedances at the higher frequencies where negative feedback is no longer feasible because of the excessive phase shift of the transistor. Other important design criteria are good input and output impedance match

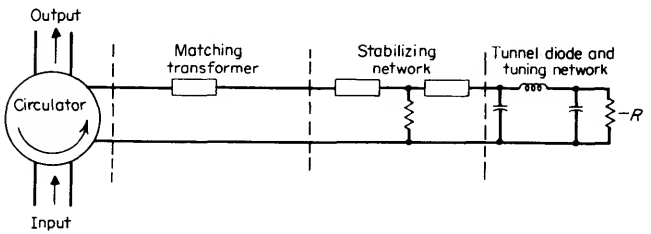


FIG 15-10 Tunnel-diode amplifier ac circuit.

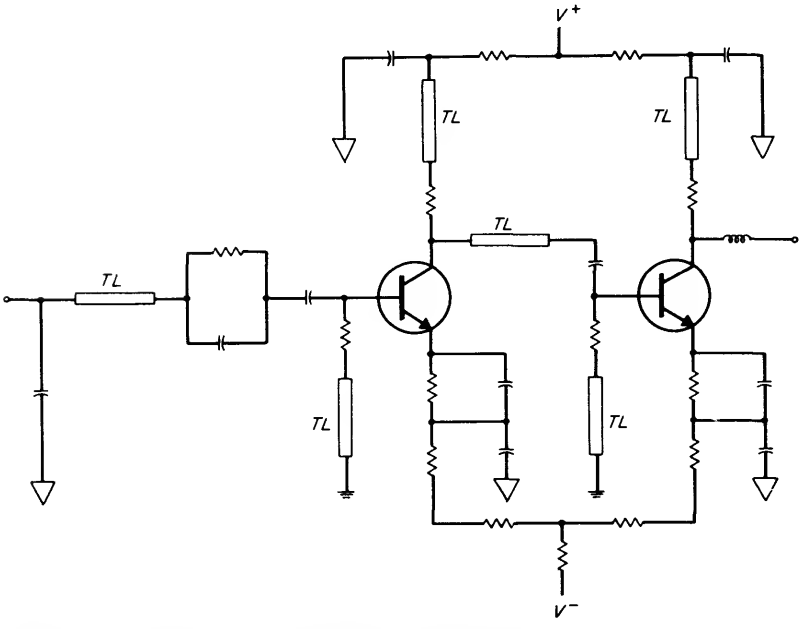


FIG 15-11 Diagram of a 0.1- to 2.0-GHz amplifier module.

and low reverse gain. The last two requirements are extremely important when modules are cascaded. The circuit configuration of the two-stage module is shown in Fig. 15-11. Although no numerical values are given, the diagram is still instructive.

Rather than transistor equivalent circuits, it is suggested that the measured two-port parameters be used directly in the design. All other circuit elements are described by some two-port parameters (y, z, or transmission parameters), and the two-port blocks are interconnected in the computerized design program in the following manner:

1. All shunt connections are made by using y-parameter matrices, $Y = Y_1 + Y_2$.
2. All series connections are made by using z-parameter matrices, $Z = Z_1 + Z_2$.
3. All cascade connections are made by using transmission parameters, $T = T_1T_2$.

The matrices can be continuously converted as the circuit is programmed step by step [14]. The quickest way to do these conversions is to have a subroutine when necessary. Once the complete circuit is modeled into the computer, the circuit is to be optimized for minimum

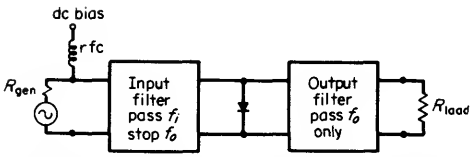


FIG 15-12 General block diagram of a frequency multiplier.

deviation from the desired gain and for best input and output impedance. Since no approximations are used, the calculated and measured results always show a strong correlation.

15-3 *Other Solid-state Microwave Sources*

The use of high-frequency transistors to generate microwave signals at low power levels has been discussed above. Microwave signals can also be generated by frequency multiplication with the use of varactors or step-recovery diodes and by new solid-state devices that exhibit negative resistance at extremely high frequencies. The last devices are the Gunn-effect oscillators and the impatt (impact avalanche transit time) diodes. After a discussion of the various methods, their operating characteristics will be compared.

Frequency Multipliers. A general block diagram of a frequency multiplier is shown in Fig. 15-12. The diode and specific circuit used depend on many things: the output frequency, the power levels at input and output, the multiplication ratio, cost and performance trade-offs, etc. Generally, except at the very highest frequencies, the diode is made to represent nonlinear reactance rather than nonlinear resistance for this application. This is because the efficiency of variable resistors falls off as $1/n^2$, whereas varactors are theoretically capable of 100 percent efficiency, providing the diode series resistance is zero [16, 17]. The term n is the output frequency of the multiplier divided by its input frequency.

There are several microwave devices that can be classified as nonlinear: (1) the varactor, (2) the step-recovery diode, and (3) the bimode diode. The first two devices are distinctly different, whereas the third is a compromise between the two design philosophies. All three of the devices have the general characteristics of Fig. 15-13.

The n layer in a typical varactor is more heavily doped than in a step-recovery diode. This results in a junction capacitance that varies with reverse voltage (Fig. 15-10). For example, in an abrupt-junction

varactor ($\gamma = \frac{1}{2}$)† the voltage across the junction is proportional to the square of the charge [18], or

$$V = kq^2 \tag{15-3-1}$$

where k = proportionality constant
q = charge in coulombs

This square-law relation gives rise to the transfer of energy from f_i to $2f_i$. For this type of device, multiplication by harmonics other than 2 is accomplished by adding *idler* circuits. These idlers allow the doubled frequency to be either redoubled or mixed with the fundamental. An idler circuit is usually defined as a resonant tank circuit, although the term is apparently sometimes meant to represent a low-impedance path for that particular Fourier component of diode voltage (resonant or not).

Higher-order frequency multiplication without idlers is possible with diodes whose γ approaches unity (full drive, not overdrive). However,

† Usually γ is defined in terms of the incremental ac elastance (reciprocal capacitance) as follows:

$$\frac{S}{S_{\max}} = \left(\frac{\phi - v}{\phi - V_B}\right)^{\gamma} \qquad \text{for } v \leq \phi$$

where S = elastance
v = diode instantaneous voltage
V_B = diode breakdown voltage
ϕ = forward junction contact potential

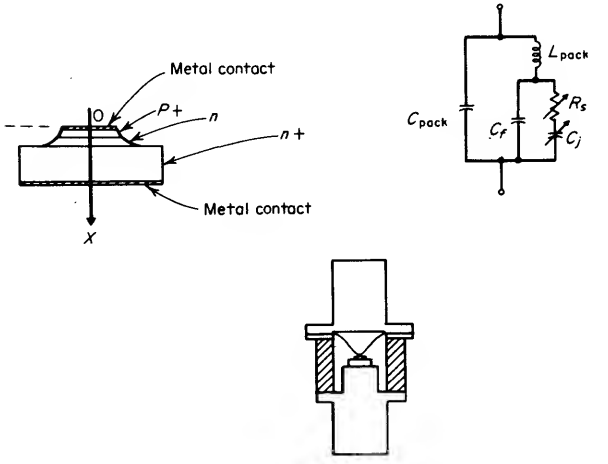


FIG 15-13 Diode construction used for most frequency-multiplier diodes whether pure varactors, step-recovery diodes, or other type.

these diodes (hyperabrupt) are very difficult to construct with high Q at very high frequencies and almost impossible at microwave frequencies. It is possible however, to generate higher harmonics at high frequencies with lower-γ varactors if they are (1) overdriven and (2) have sufficiently fast recovery times from their forward charge storage state so that they behave much as if they were varactors with higher γ.

It has been shown [19] that if this forward charge storage effect (overdrive) is to be used, the fastest recovery is achieved by using the profile of the step-recovery diode, which has γ approaching zero. For the step-recovery diode, overdrive must be used since no nonlinearity occurs until the diode is driven hard into conduction.

The n-layer in a step-recovery diode is so lightly doped (Fig. 15-10a and b) that it is usually called an i layer (intrinsic). This light doping means that all the charge is swept out quickly and completely on application of a low negative bias voltage, and the diode capacitance is ideally constant from this voltage (the "punch through" voltage) out to the breakdown.

The diode will behave as a large storage capacitor under forward storage if the period of the driving waveform is short compared with the effective minority carrier lifetime τ, that is,

$$\frac{1}{f_i} \ll \tau, \text{ where } f_i \text{ is input frequency} \qquad (15\text{-}3\text{-}2)$$

When all the stored charge has been removed by negative current, the diode should ideally "snap" back to the depletion capacitance value in zero time. If this transition time is much shorter than the period of the highest frequency, a very simple diode model results. That is, if the transition time $t_t < 1/f_o$, then the diode may be viewed, as shown in Fig. 15-13, as a two-state switching capacitor controlled by the sign of the charge.

Circuits for Multipliers. Figure 15-14 shows a typical varactor doubler circuit. Element values are chosen by using calculated values for the circuit [18, 20, 21]. The inductances L_i and L_o are chosen to tune with the average reactance at the input and output frequency. The input and output filters provide transformation of the generator and load impedances to the values as determined in a complete optimization procedure [20]. The filters must then separate input and output signal paths so that no input signal appears at the output port and vice versa. Low-pass or bandpass matching techniques may be used [22, 23]. Bandpass matching techniques may be preferable when R_i and R are not significantly different from R_g and R_L, since in a low-pass filter matching network, signal separation depends upon these relative values.

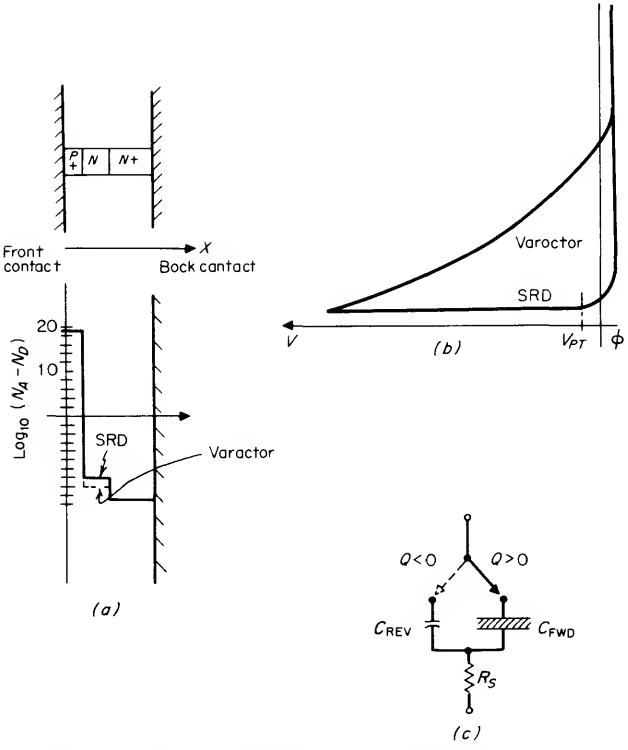


FIG 15-14 Varactor and step-recovery diode profile and CV characteristics: (*a*) doping profiles, (*b*) CV characteristics, and (*c*) switching capacitance equivalent of step-recovery diode. The equivalent circuit of the step recovery is a simple charge-controlled switch in the frequency range of $1/t < f < 1/t_\tau$.

The optimum power output and efficiency η of varactor multipliers have been calculated [20, 21]. They are related to two defined analytical parameters α and β by

$$\eta = \exp\left(-\alpha \frac{f_o}{f_c}\right) \tag{15-3-3}$$

$$P_o = \beta \omega_i C_{\min}(\phi - V_{B\tau})^2 \tag{15-3-4}$$

where f_o = output frequency
f_c = cutoff frequency = $(2C_{\min}R_s)^{-1}$
$C_{\min}$ = capacitance on reverse bias
R_s = series resistance of diode
$V_{B\tau}$ = reverse breakdown voltage

Equation (15-3-4) shows that for the power, output increases directly with the diode capacitance. As the capacitance is increased, however, the input and output impedances decrease, since they are given by

$$R_i = R_s A \frac{\omega_c}{\omega_i} \tag{15-3-5}$$

$$R_o = R_s B \frac{\omega_c}{\omega_i} \tag{15-3-6}$$

Values for α, β, A, and B may be found in Burckhardt [20].

When R_i gets so low that it is equal to the loss resistance (owing to finite Q_u of the filter elements and the diode series resistance), no further gain in power output can be made by increasing $C_{\min}$. Frequently, other limiting factors, such as overheating and $V_{Br} - t_t$ trade-offs, are encountered before this limit is reached.

The varactor multiplier shown in Fig. 15-14 has the advantages of:

1. Providing high output power with good efficiency
2. Being calculable
3. Being broadbandable as a doubler

Its main disadvantages are that:

1. It is primarily a low-order multiplier.
2. It requires retuning at different input power levels.
3. For other than doublers the circuit becomes extremely complicated because of the idler circuits.
4. The broadbanding of doublers of this type is feasible, but the broadbanding of idler-dependent multipliers is quite difficult because of the idler interaction with circuits that determine f_i and f_o.

Figure 15-15 shows a step-recovery multiplier circuit. Both lumped circuits and distributed circuits can be used with the step-recovery diode to form frequency multipliers [24, 25]. The operation of this type of multiplier is briefly as follows:

1. The diode is driven hard into forward charge storage and remains conducting for most, $[1 - (1/2n)] \times 100$ percent, of the input period.
2. Just prior to the short period when the diode is off, most of the stored energy is in the inductor L_i since the charge in the diode, q, is approaching zero.
3. When the diode goes to zero charge, it changes to a high-impedance state and the energy in L_i transfers to the reverse bias capacitance of the diode as a large pulse of reverse voltage, limited to less than V_{Br} in amplitude.
4. This pulse shock-excites the output filter, which now rings freely at the output frequency loaded on the diode end by the forward series resistance of the diode (it starts conducting again immediately after pulse) and by the load on the other end.

Experimentally and theoretically it has been found [24, 25] that the pulse, t_p, should be one-half to one cycle wide at f_o. This width is controlled by L_i through

$$t_p = \pi \sqrt{L_i C_{\min}} \tag{15-3-7}$$

where L_i = drive inductance
$C_{\min}$ = diode reverse bias capacitance

$$t_p = \frac{x}{2f_o} \qquad 1 < x < 2$$

The capacitance C_T resonates with L_i at f_i and should itself be nonresonant up to at least the frequency of the highest Fourier components of the diode voltage (harmonics of f_i appear in the diode voltage up to $f = 3/2t_p$). The input resistance across the C_T terminals is known to be approximately equal to the inductive reactance $\omega_i L_i$, which value can be used to design the input matching network. The output inductor is approximately equal to L_i for optimum energy transfer from input to output circuits. The output filter's first element can be either a series or shunt capacitor which resonates with L_i at f_o. A popular output resonator is shown in Fig. 15-15*b*. Ideally, an impulse traveling down the line reflects from C and returns to the diode just as it is closing again. It then resonates in this shorted quarter-wave link until the energy in the pulse has been dissipated in the load to the right of C.

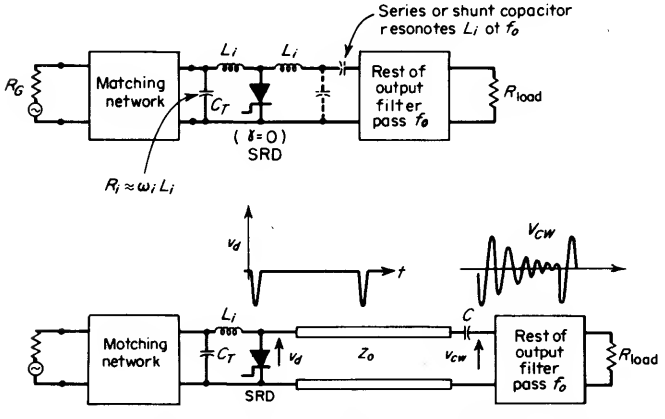


FIG 15-15 Two generally used step-recovery multiplier circuits. The diode voltage is a narrow pulse with Fourier components at the desired output frequency $\omega_o = n\omega_i$. This pulse shock-excites the output filter once per input cycle.

The output filter of Fig. 15-15 may be designed around bandpass impedance matching techniques [22, chap. 11], once the equivalent output resistance of the step-recovery diode is known. This may be determined experimentally by using a one-element filter as in Fig. 15-16.

The capacitor C (the output coupling capacitor) and ℓ are adjusted to give maximum power output in the line, P_n. This corresponds to saying that the diode is seeing the best impedance to accomplish this. The value of Z_0 should be large enough so as not to short out the impulse during its formation, $Z_0 > 1/\omega_o C_{\min}$. Transforming back through Z_0, one can calculate R_o, the equivalent output impedance, to use in the design output-filter network. In designing the bandpass impedance network, the formation of the pulse across the step-recovery diode must again be kept in mind. The transient impedance of the first elements of the output filter must not short out the pulse. The series inductor of Fig. 15-15*a* and the quarter-wave line of Fig. 15-15*b* are examples which work and do not tend to short out the pulse. After these elements are included, the remainder of the output-filter design is conventional [22, chap. 11].

Alternatively, an equivalent dc-forward-bias point which will simulate the effective output impedance of the diode can be established for later experimental filter adjustment as shown in Fig. 15-17. With no rf input to the input port, the diode is forward dc biased to several milliamperes and the swept output impedance (in the output frequency band at the output port) is measured with a reflectometer technique. At some bias point the output will be nearly matched at $nf_i = f_o$. This bias point and swept technique can then be used again after the entire output filter has been designed.

Bias Circuits. The varactor multiplier actually is often driven into a condition of forward charge in order to achieve a larger change in reactance. The step-recovery multiplier always is driven forward to obtain charge storage, and some recombination current flows. This recombination current can be used to develop negative dc-bias voltage across a

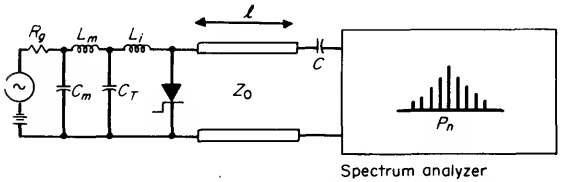


FIG 15-16 Measurement setup that can be used to determine the equivalent output impedance R_o of the step-recovery diode. Knowing Q_L (spectrum analysis), Z_o, and the fact that P_n is maximum allows calculation of R_o.

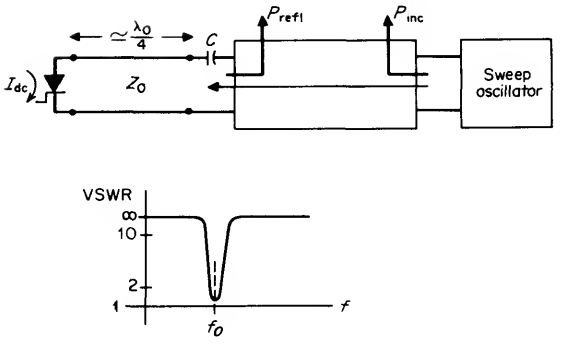


FIG 15-17 An approximately equivalent output impedance can be achieved by using dc forward bias of a few milliamperes. Swept-reflectometer techniques on the output will show vswr response of the output filter.

bypassed resistor (Fig. 15-18). In the case of the step-recovery diode, it can be shown that the bias voltage required to stay "in mode" increases directly with the input voltage, and that the bias voltage developed across R_b also is linear with E_i. Then the step-recovery diode multiplier, at least theoretically, should have a linear P_i versus P_o curve between *threshold* and *saturation*. This is in contrast to the varactor multiplier, which must be tuned at each P_i.

The self-bias feature has another advantage in that temperature compensation, which holds the bias constant over temperature, is accomplished by matching the temperature coefficient of R_b with the temperature coefficient of τ, the effective minority carrier lifetime of the step-recovery diode (which is approximately linear at 0.7 percent/°C).

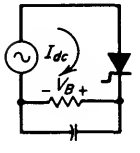


FIG 15-18 The very small amount of rectification that does occur in the step-recovery or overdriven varacter can be used to develop self-bias.

The main advantages of the step-recovery diode multiplier circuit are that:

1. The circuitry, even for high-frequency ratios, is simpler than for the varactor multiplier.
2. Once set up and in mode, the self-biased step-recovery diode multiplier has a linear characteristic of power input to power output (an important feature in a large system that has realistic power-level tolerances over temperature and frequency).

3. The design of a step-recovery diode multiplier can be divided into several parts (impulse generator, damped-waveform generation, bandpass filter, overall multiplier) which can be evaluated independently along the way. This is an important practical consideration.

4. Circuits of the step-recovery diode are easier to tune because there are fewer stages and all are less sensitive to power level.

The main disadvantages are:

1. Bandwidths achievable are smaller because of higher ratios used. The maximum bandwidth for any single-stage n-fold multiplier with perfect filter is $BW \approx f_i$.

2. Output frequencies are generally limited to less than about 20 GHz because of transition times achievable.

3. A definite relation between V_{Br} and transition time limits the available power output. For example, in a $\times$ 5 multiplier, $P_{o\,max} \approx 10$ W at $f_o = 2$ GHz and 1 W at 10 GHz.

Stability of Frequency Multipliers. When cascaded or driven by an imperfect driving circuit, frequency multipliers often break into various kinds of spurious oscillations with variation in temperature, power level, or frequency. This type of instability has historically been the biggest problem in the practical implementation of frequency multipliers.

Stability is mainly achieved by good clean circuitry with short interconnecting lengths, nonresonant elements, and no series resonances in the circuit at $f < f_i$. Whenever instabilities of a comblike-line nature appear, the generation of parametric oscillations at $f = f_{sr}$ should be suspected, where f_{sr} is the resonant frequency of an inadvertent series resonant circuit. The explanation is simple: If there is a resonance at f_{sr} and the diode is "pumped" at f_i, negative resistance *can* appear at $f_i - f_{sr}$.

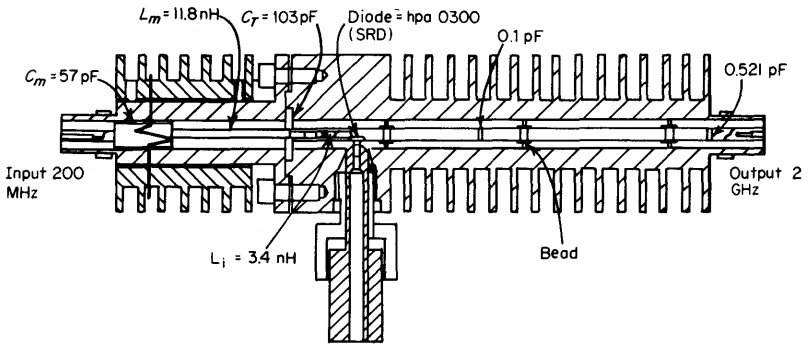


FIG 15-19 Cross-sectional drawing of a $\times$ 10 (200 to 2,000 MHz) step-recovery diode frequency multiplier. The tuning bandwidth of this multiplier is 100 MHz (to −1-dB point) and P_o = 2.5 W.

If it is larger than the circuit resistance, oscillations will occur. Elimination of all parasitic series resonances has led to multipliers that were stable even when terminated by a short circuit at any phase.

Figure 15-19 shows a cross-sectional drawing and photograph of a × 10 frequency multiplier with step-recovery diode, which uses the design described above. It achieves linear power output versus input, an output that is flat within 1 dB over a band of 100 MHz, and a power output of 2.5 W with efficiency of 17 percent.

15-4 *Solid-state Microwave Oscillators* [26]

The Gunn oscillator and the impact-avalanche transit-time (impatt) diode are two-terminal semiconductor devices that oscillate when operated properly in a tuned cavity.†

The Gunn oscillator operates as a result of the properties of gallium arsenide and other III-V semiconductor compounds. Impatt diodes are p-n junctions operated in reverse breakdown. The avalanche process in the impatt generates an effective average negative conductance that can be used for amplification or generation of oscillations. There are several modes of operation for both devices, but to date the Gunn or bulk-effect devices are somewhat less noisy and more broadly tunable, whereas the impatts provide higher average power output and are useful to higher frequencies.

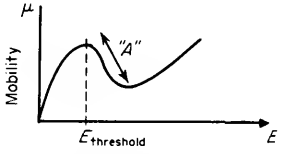


FIG 15-20 The mobility, and hence velocity, of carriers in GaAs is a function of the applied E field. A region of negative differential mobility, A, corresponds to a region of negative differential resistance.

Gunn Oscillator. The Gunn-effect device is not a diode; rather, it is a piece of bulk semiconductor material with front and back metal or ohmic contacts. Gallium arsenide is most commonly used, although some other intermetallic (III-V) compounds are sometimes used. These compounds have been found to have a region of negative differential mobility above a certain threshold electric field (3,000 V/cm) but below avalanche breakdown, as shown in Fig. 5-20. Current flow is proportional to the mobility, $J = en_0\mu E$, and so a negative differential resistance exists when dJ/dE is negative

$$\frac{dJ}{dE} = en_0 \frac{d\mu}{dE} < 0 \qquad (15\text{-}4\text{-}1)$$

† Actually, the basis for the diode that exhibits negative resistance was made by W. T. Read, Jr., in his 1958 paper [32]. J. B. Gunn reported microwave oscillations in bulk III–V semiconductors in his paper entitled "Microwave Oscillations of Current in III–V Semiconductors" (*Solid State Communications*, vol. 1, p. 88, September, 1963).

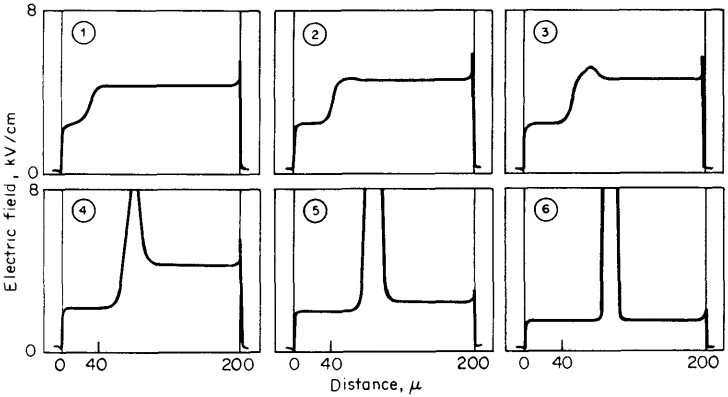


FIG 15-21 Sequence showing development of domain in 200-μm sample of GaAs. The domain was initiated by a small doping change near the cathode end of the sample.

where n_0 is the background doping density. Negative resistance is shown in region A of Fig. 15-20. When biased to this region, the diode oscillates when tuned and loaded correctly.

Several modes of oscillation are useful:

1. The domain mode
2. The quenched-domain, or hybrid, mode
3. The limited-space-charge-accumulation (LSA) mode

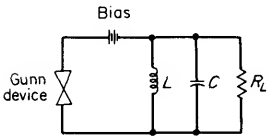


FIG 15-22 Gunn device in an rf circuit. The rf voltage across the tank circuit can force the terminal voltage below the value required to sustain the domain during its transit quenched-domain mode. At higher frequencies the domain may not even form (LSA mode).

The domain mode is illustrated in Fig. 15-21. When the field at the anode builds up higher than $E_{\text{threshold}}$, an accumulation of mobile space charge begins to form and propagate toward the cathode across the sample in a packet, or *domain*. When the product of the length l and the available charges n_0 is large enough, this domain will reach an amplitude such that most of the device voltage is across the accumulation layer and the field elsewhere is below threshold. When the device is placed in a microwave cavity, as in Fig. 15-22, its voltage is controlled by the cavity resonance, and the cavity is excited into oscillations by the periodic pulses from the device. The output frequency is controlled primarily by the length of the sample, the domain velocity, and the resonant frequency and Q of the external cavity. Tuning ranges greater than one octave have been achieved in the domain mode.

When the cavity resonance frequency is substantially higher than the transit time frequency, the space charge wave may be *quenched* because the terminal voltage drops below that required to sustain the space charge layer for the complete transit. This mode, the quenched-domain mode, allows the device to operate at frequencies significantly above the transit time frequency [27].

When the frequency has become so high that the domain never has the chance to form, we have the LSA mode [28]. In this mode, since there is no voltage drop across a charge layer, the entire length of the device has effective negative differential resistance.

In the domain mode and the quenched-domain mode, it has been shown that Pf^2Z = constant [29], that is, power output varies inversely as the square of the operating frequency. This is not so for the LSA mode. The limits on the LSA mode are often thermal. Gallium arsenide has high thermal resistance that leads to overheating, which can impair the negative mobility mechanisms of the device. Therefore, LSA oscillators are often pulsed, rather than continuous-wave, devices.

Broadly tunable oscillators with use of the Gunn effect have been built in several laboratories. For example, a Hewlett-Packard Company device has been tuned over a 4- to 12-GHz range with a yig resonator [30].

Power outputs for continuous-wave Gunn-effect devices are generally not as high as for avalanche devices because of the thermal impedance problems of gallium arsenide. Continuous-wave oscillations of several hundred milliwatts in the 4- to 20-GHz range appear to be reasonable, while pulsed operation in LSA modes have great potential even at 50 to 100 GHz.

Noise performance of Gunn oscillators, close to the carrier, appears to be generally some 20 dB better than the silicon avalanche-diode oscillators, but they are still somewhat noisier than transistors. Present AM and FM noise performance of Gunn-effect devices and avalanche devices are summarized in Fig. 15-23 [31].

Impatt Diode Oscillators. The impatt diode generates an effective negative resistance by the combination of charge generation by avalanching and the time delay of the transit of the charge across the nonavalanching portion of the diode. The negative resistance can be generated in p-n junctions and p-i-n devices, but is most easily visualized in the Read diode structure of Fig. 15-24 [32].

This structure consists of an n-p junction at the cathode end and a p region at the anode end, separated by an intrinsic region (i layer) that is the drift space. When reverse biased, the n-p junction tends to break down first and form a pulse of charge where the field is highest. When driven far enough above $E_{critical}$, this pulse of charge becomes very narrow and approaches an impulse at high ac levels (Fig. 15-25). The peak of the charge pulse lags the peak of the ac voltage which generated it by 90°.

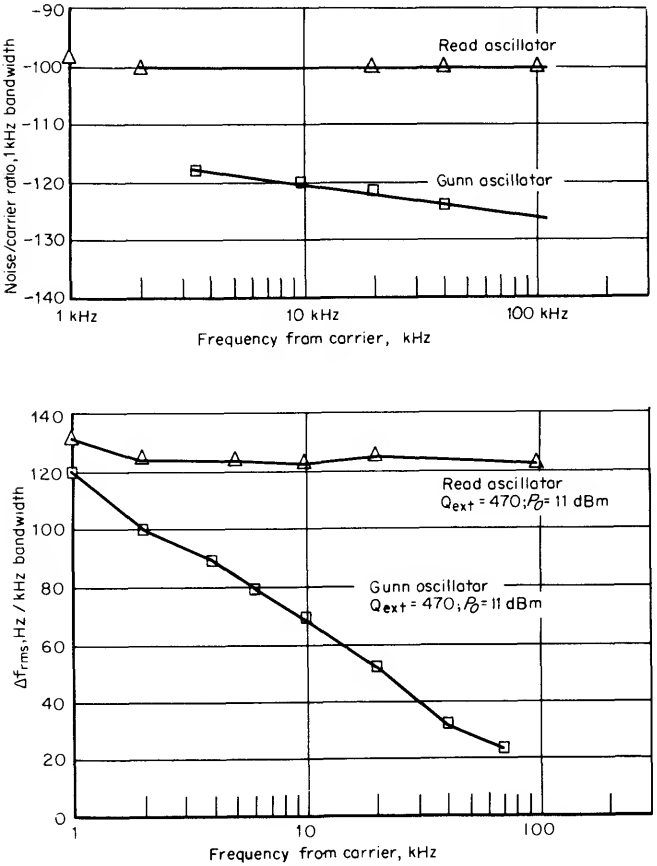


FIG 15-23 Amplitude- and frequency-modulation noise spectra of Read and Gunn oscillators. ***(From Ref. 31, by permission of the authors.***

This pulse of charge enters the drift space (width W) where the E field is high enough to maintain saturated velocity of the carriers (10^7 cm/sec). The particle current flowing in the external circuit while the charge transits the drift space is equal to the average current in the drift space. Since the velocity is constant, the current (at the contacts) is constant from the time the charge enters until it leaves the drift space. If the drift-space width is $W = \frac{1}{2}f_0$, the generated square pulse of current will be exactly 180° out of phase with the ac voltage that generated it. Idealized current, voltage, and charge waveforms are shown in Fig. 15-25

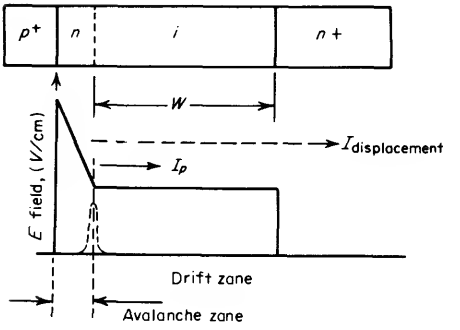


FIG 15-24 Read structure and static field distribution.

for a large ac signal. For smaller ac signals, the charge waveform is somewhat less like an impulse, or somewhat broader. This tends to give rounded leading and trailing edges to the current pulse, and this in turn contributes some positive resistance to be averaged over the cycle. At even the lowest signal levels (noise level, for example) there is a time average negative resistance. Once this has been established, the growth of continuous-wave oscillations follows.

Avalanche Oscillators. Most practical avalanche transit-time oscillators use silicon, germanium, or gallium arsenide p-n or p-i-n diodes because they are so much easier to fabricate and because they accomplish essentially the same thing as the more complicated Read structure. Silicon has the lowest thermal resistance of these materials, and its process technology is further advanced. Gallium arsenide is primarily attractive because of its potentially better noise properties. Its chief

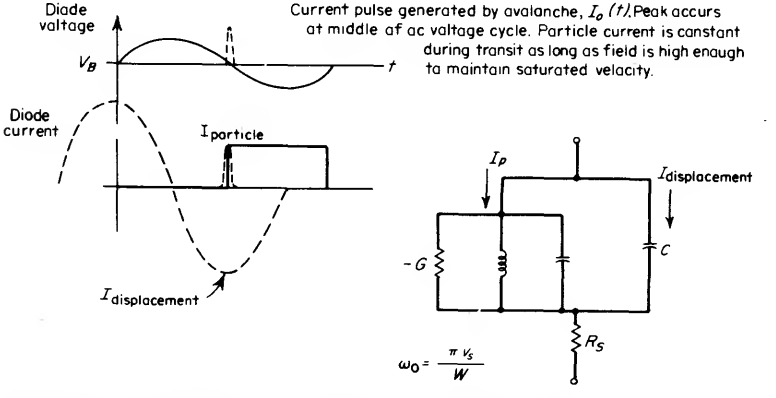


FIG 15-25 An ac cycle in a Read diode and a circuit that represents the first-order effects.

drawbacks are its high thermal resistance and the status of gallium arsenide technology.

The maximum power output of impatt oscillators is limited by thermal considerations. Enormous power densities exist at the junction of the device. For example, a typical x-band diode might have an area of 1.5×10^{-4} cm^2, a high-current breakdown of 100 V, and a dc-bias current of 150 mA to produce 1 W of rf output. This corresponds to a current density of 1,000 A/cm^2 and a power density of 10^5 W/cm^2. The state-of-the-art data of Fig. 15-26 were obtained at Bell Telephone Laboratories by using diamond heat sinks, which have 10-times better thermal conductivity than oxygen-free copper.

The negative resistance of the avalanche device exists over a fairly wide frequency range. For example, the circuit shown in Fig. 15-27*a* mechanically tunes a single diode from 6 to 10 GHz with 200- to 300-mW output.

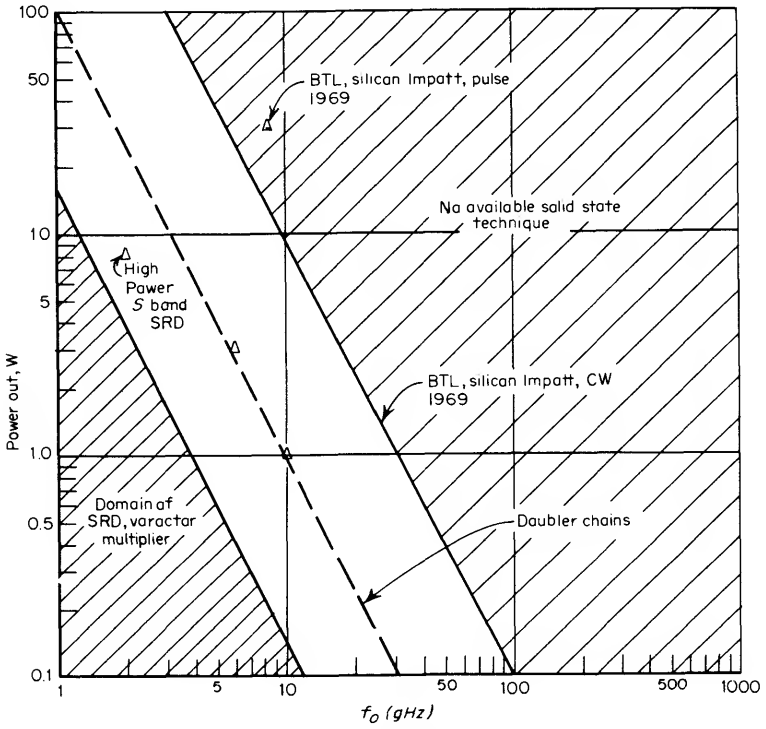


FIG 15-26 Continuous-wave power output capability of various solid-state sources.

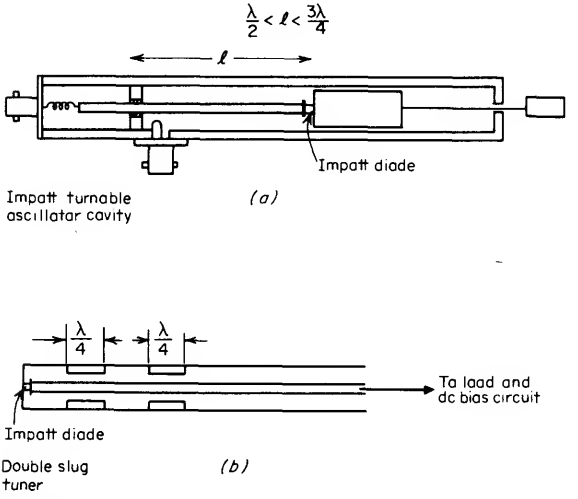


FIG 15-27 Two impatt-oscillator circuits: (*a*) impatt-tunable oscillator cavity, (*b*) double slug tuner.

Another useful tunable oscillator circuit is shown in Fig. 15-27*b*. The two slugs form quarter-wave lines of low impedance that transform the load impedance to the optimum value for the diode. Typical avalanche diodes studied by this writer operate best with a very low resistance (about 1 Ω) in series with an inductive reactance of 20 to 30 Ω. In a package these values are altered by the parasitic inductance and package capacitance.

The FM noise performance of an impatt oscillator depends on the Q of the oscillator cavity. High-Q cavities lead to lower FM noise, as can be seen in the expression for FM noise [22]

$$\Delta f_{ms}^2 = \frac{f_0^2}{Q_{ext}^2} \frac{kT_0BM}{P_0} \tag{15-4-2}$$

$$M = \frac{T_{eff}}{T_0} \tag{15-4-3}$$

Frequency-modulation noise close to the carrier can be improved by phase locking. The quieter the free-running oscillator, the lower the permissible locking-signal level is and the further from the carrier the quieting will have effect. Sufficiently far from the carrier, the noise level approaches the free-running oscillator noise.

Amplitude-modulation noise in impatt oscillators is strongly dependent on the bias circuit [23]. When the diode is terminated improperly, para-

metric coherent sidebands close to the carrier can be generated at levels comparable to the main signal. When the bias circuit is correctly designed, the AM noise becomes a function of the Q of the oscillator cavity, the power output, and M, for a single sideband,

$$\frac{P_{\text{AM noise}}}{P_{\text{carrier}}} = \frac{4kT_0BM/P_0}{S^2 + \left(\frac{2Q_{\text{ext}}f_m}{f_0}\right)^2} \tag{15-4-4}$$

At frequencies further away from the carrier than f_0/Q_{ext}, the AM noise falls off at the rate of 6 dB per octave.

Typical AM and FM noise data for free-running impatt oscillators are given in Fig. 15-23.

15-5 Comparison of Solid-state Sources

Oscillators have several important parameters: power output, tunability, long- and short-term stability, AM and FM noise, and capability for modulation.

Continuous-wave power output capabilities of each of the sources is illustrated in Fig. 15-26. The impatt diode is seen to have a commanding edge in this respect. All the devices, except the LSA-mode, bulk-effect device, have a power output that decreases as $1/f^2$. This product of power and reactance was discussed in detail by Early and later by Johnson and DeLoach [33] for various devices, including transistors, varactors, Gunn devices, and avalanche diodes. The LSA-mode, bulk-effect device is the single exception that does not follow the $1/f^2$ law. For pulse operation, with low average power, it shows the largest power output of the group. The power output of varactor doubler chains, with balanced doublers, is competitive with the most powerful continuous-wave Gunn-effect devices. A doubler chain of this type requires several stages to achieve the desired output frequency, and at each stage a very high power device (for that frequency) is required. The power output of a $\times$ 5 step-recovery diode multiplier is somewhat lower than the most advanced varactor doubler. However, fewer stages and less complicated circuit interaction are involved with the step-recovery diode multiplier.

Mechanical or electrical tunability, or both, of oscillators is an important feature in many systems. Even a small amount of tunability may make the difference in whether a parametric amplifier performs or not, or whether a low-noise, phase-locked FM system is achievable. Certainly the most tunable of the group is the low-power Gunn-effect oscillator. Devices of this type that are yig-tuned have been built to tune with 40 mW of output power from 4 to 13 GHz [30]. Impatt oscillators

are second in this class. Mechanical tunability over an octave is achievable with a power output of 200 to 300 mW of power output. Electronic tuning of impatt oscillators with varactors is strongly dependent on the power output and tuning range required. Broader tuning is achieved with the lower output powers. Electronic tuning over a range of 200 MHz has been demonstrated at 16 GHz and $P_0 = 50$ mW [34] in a microwave integrated circuit. Step-recovery multiplier chains are limited practically to a fractional tuning bandwidth (referred to the output frequency) of $1/2n$ percent, where n is the largest harmonic number in the chain.

To achieve good long-term frequency stability, one wants somehow to lock the microwave signal to a reference crystal control oscillator. In a straight multiplier chain, if stability to better than $1/10^3$ is required, the fundamental oscillator is crystal locked, which automatically establishes the long-term stability of the microwave signal. (This technique, however, degrades the ratio of FM noise to carrier, as discussed below.) In order to achieve long-term stability from the microwave oscillator devices, some form of phase locking is used. The simplest phase-locking scheme is shown in Fig. 15-28. A low-level reference signal is injected into the oscillator cavity, the presence of this signal tends to "lock" the differential phase of the two signals to some value, and the frequencies are made identical. An alternative phase-locking technique uses a sampler as in Fig. 15-29. A small time segment of the rf voltage is "sampled" onto a holding capacitor once per cycle of the low-frequency (50- to 100-MHz) reference oscillator. Any drift from sample to sample in the sample voltage is amplified and used to correct the microwave oscillator frequency (through a varactor) until locking is achieved.

Properties in AM and FM noise for free-running impatt oscillators and Gunn-effect devices are shown in Fig. 15-23. The noise is shown as a function of f_m, the distance of the measuring slot away from the carrier. The significance of local oscillator noise is illustrated in an AM example in Fig. 15-30. A single-ended mixer and an intermediate frequency close to the carrier were chosen deliberately to illustrate the point that local oscillator noise could cause a degradation of 14 dB in detectable signal

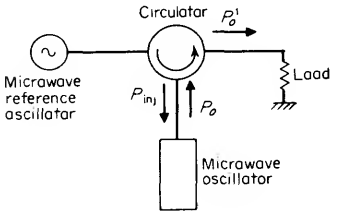


FIG 15-28 An injection phase-locking circuit using a circulator.

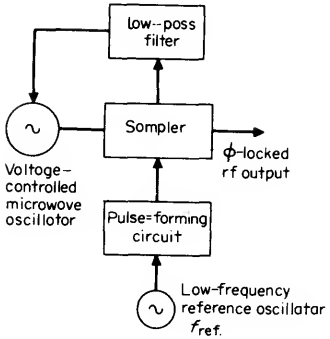
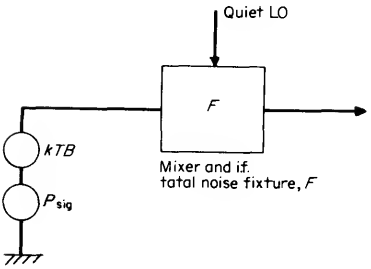
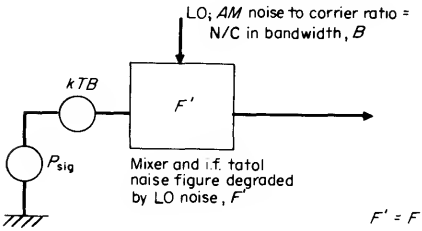


FIG 15-29 Phase-lock loop using sampler as a phase detector.





$$F' = F + \frac{\text{LO noise in bandwidth, } B}{kTB}$$

$$\frac{F'}{F} = 1 + \frac{N_{LO}}{FkTB} = 1 + \frac{P_{LO}}{FkTB}\left(\frac{N}{C}\right)$$

Since

For $P_{LO} = 1$ mW $B = 1$ kHz
$kTB = 10^{-14.4}$ (−144 dBm)
$N/C = 10^{-12}$ (−120 dB)
$F = 10$ dB

$\frac{F'}{F} \approx$ 14 dB degrodotion

FIG 15-30 Degradation in apparent noise figure to local oscillator noise.

level. (Amplitude-modulation systems generally use balanced mixers and higher intermediate frequencies to avoid this kind of problem.) The problem in FM systems is similar in that incoherent phase variation in the information band masks the desired signal at some low level. In FM systems balanced mixing does not cancel out local oscillator noise. However, phase locking to a quiet reference source will improve FM noise performance close to the carrier. Impatt devices have been generally found to yield flat AM and FM noise spectra. The noise of the Gunn oscillator decreases with increasing f_m. This decreasing noise makes the Gunn device attractive for local oscillator use in applications such as that described above where there is a 10- to 60-MHz intermediate frequency and possibly a balanced mixer.

The AM and FM noise properties of multipliers are somewhat more complicated. A stable multiplier chain has a characteristically well-behaved AM noise. The AM noise tends to decrease with increasing f_m regardless of the multiplication ratio. The FM noise deviation is linear with n; the multiplication ratio is

$$\Delta f = n\,\Delta f_{\text{drive}} \tag{15-5-1}$$

where Δf = FM deviation at output
Δf_{drive} = FM deviation at input
Expressed in noise power, the ratio of single sideband to carrier is degraded by the factor n^2,

$$\frac{P_n}{P_o} = \left(\frac{n\,\Delta f_{\text{drive}}}{2f_m}\right)^2 \tag{15-5-2}$$

As a result, to attain the lowest Δf, we should like to use the lowest possible n. This of course conflicts with the desire to start with a stable low-frequency crystal-controlled oscillator and just to multiply by n up to the desired f_o. A good solution to this problem is used in several commercial solid-state sources, as shown in Fig. 15-31. A fundamental transistor oscillator in high L band is phase locked to a low-power, high-order multiplier output from a 50- to 150-MHz crystal oscillator. The

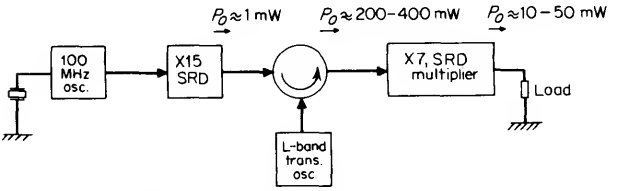


FIG 15-31 Block diagram of a low-noise, stable solid-state source.

L-band output, which is now crystal stable, is multiplied by a much smaller n to the desired output frequency.

Modulation of sources falls into three principal groups: (1) AM, (2) FM, and (3) pulse modulation. Sine-wave AM is simple only with the self-biased step-recovery multiplier. (Even then the AM must have a small modulation index and the multiplier must be in its linear power range.) The step-recovery diode can also be frequency modulated by varying the bias voltage. Small bias changes cause the recovery angle to change by $\Delta\phi$, and this changes the output phase by $n\,\Delta\phi$. Phased array steering by use of this property has been proposed. Varactor multipliers tend to be hard to frequency or amplitude modulate independently, although FM of the driver is possible and has been demonstrated. Impatt-diode oscillators and Gunn-effect oscillators may be tuned or modulated by use of yig or varactor tuning elements. Pulse modulation may be accomplished by pulsing the bias to the diode or by use of a p-i-n modulator on the output of the continuous-wave device. Programmed properly, the latter may also be used for sine-wave modulation.

15-6 Microwave Signal Generators

This section treats generators of standard microwave signals with frequencies above 2 MHz. The major characteristics of these signals are specified and calibrated; the specifications can cover tuning ranges, amplitude, type and degree of available modulation, output impedance, and the accuracy and resolution of these quantities. Also specified are spurious outputs such as distortion, hum, and noise. It is important for the engineer to understand the various configurations of standard-signal generators, so that he can compare and use them satisfactorily.

Block Diagram. The generator that starts with a single stable frequency and synthesizes a wide range of frequencies by multiplication and division is discussed in Chap. 6. Below we discuss the direct signal generator, whose output frequency is the same as its oscillator frequency. We also briefly discuss the heterodyne generator.

Figure 15-32 is a block diagram of the direct generator, and Fig. 15-33

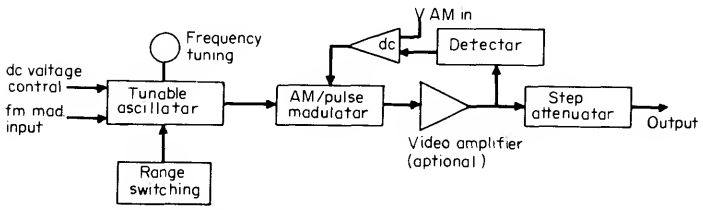


FIG 15-32 Simplified block diagram of a direct signal generator.

FIG 15-33 Direct signal generator, Hewlett-Packard 8616A, 1.8 to 4.5 GHz.

shows a photograph of such an instrument. The desired output frequency is formed in a single tunable oscillator usually covering a range of at least 2:1 in frequency continuously, with provision for switching the range-determining elements to increase the total frequency range covered by the instrument. If the generator is to provide frequency modulation and the basic oscillator is of the mechanically tuned variety, a voltage- or current-controlled reactance will be provided in the oscillator circuit. The oscillator is followed by signal-conditioning circuits such as amplitude or pulse modulators, amplifiers, and attenuators which further determine the output characteristics of the generator and also serve to isolate the basic oscillator from the effects of variations in load.

Figure 15-34 is a diagram of the heterodyne generator, usually associated with instruments that automatically sweep through a range of frequencies. However, sweepers usually have a continuous-wave mode of operation. In Fig. 15-34, the output frequency is the difference frequency between a fixed oscillator and a tunable oscillator. Obviously, wide range on a single band is possible, and the user may consider this

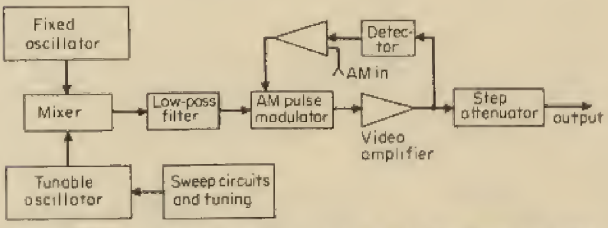


FIG 15-34 Simplified block diagram of a heterodyne signal generator.

feature a good one, but there are other specifications that may suffer. For example, in Fig. 15-34 let the tunable oscillator tune an octave from 2.0 to 4.0 GHz and the fixed oscillator be set at 1.9 GHz; then the generator will produce an output signal from 100 MHz to 1.8 GHz (not 2.1 GHz since the low-pass filter must remove the 1.9-GHz signal from the output). The accuracy and settability of frequency are now determined by how well the tuned oscillator can be set to a given much higher frequency than that desired and how accurately the fixed and variable oscillators have been factory calibrated. Additionally the output noise and stability of the generator are now determined by the fluctuations of two oscillators rather than one as in a direct generator. Heterodyne generators will be discussed below, under the topic of sweep generators.

Oscillator Stability. In a precision signal generator, the accuracy, stability, and settability of both frequency and amplitude are determined largely by the quality of the oscillator in the instrument. A survey of microwave oscillators has been given above, but the extreme difficulty in designing oscillators with adequate electrical and mechanical stability is emphasized here. For instance, the magnetic gap in a yig tuner is typically about 0.05 in., and resonant frequency is proportional to flux. Therefore, microinches of motion can cause frequency deviations of the order of megahertz. Microphonic FM is a severe problem.

If the frequency or phase variations in an oscillator caused by line frequency effects can be ignored, then they are random and can be treated by means of the principles in Chap. 4. The power spectral density of the phase variations (symmetrical about the carrier) is a valuable indication of oscillator quality, and the curve is easily measured. If ω_m is the phase deviation, and $S_\phi(\omega_m)$ is the power spectral density in radians squared per hertz as a function of ω_m, then Fig. 15-35 shows a typical curve for a good solid-state oscillator. In the low-frequency region where the slope of the asymptote is 9 dB per octave, phase noise is determined mainly by parameter fluctuation, which has a $1/f$ power spectrum. White noise in the oscillator feedback circuit is predominant in the center region, and the noise in the active device persists after $\omega_0/2Q$, where ω_0 is the oscil-

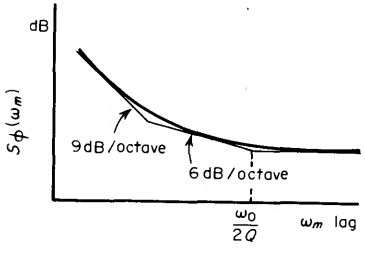


FIG 15-35 Power spectral density of an oscillator.

lator center frequency and Q is the quality factor for the resonant tuning device. It is obvious that the loaded Q is a critical design factor in a precision oscillator with low residual FM. Since the spectral density of the *frequency* fluctuations is

$$S_f(\omega_m) = \frac{dS_\phi}{d\omega_m} = \omega_m^2 S_\phi(\omega_m) \tag{15-6-1}$$

a curve similar to that in Fig. 15-34 can be plotted with the frequency spectral density as ordinate [35].

Signal Generator Specifications. The following paragraphs, although far from complete because of space limitations, will attempt to indicate acceptable and currently available specification limits for a standard-signal generator. Most of the specifications also apply to frequency synthesizers, the principles of which were discussed in Chap. 6. The material in Chap. 10 is also germane.

Output Frequency Characteristics

Range. Standard-signal generators usually have multioctave frequency coverage where the upper frequency limit is less than 1 GHz. Above 1 GHz, because of the difficulty in obtaining suitable active devices and in switching the range of the tuning mechanism, the standard-signal generators usually only provide slightly more than an octave of frequency coverage, and above 8 GHz the range is usually less than an octave.

Resolution and Accuracy. Resolution is commonly quoted in terms of frequency and refers to the minor calibration marks of the frequency readout. It should not be confused with accuracy of calibration since it only serves to indicate the magnitude of relative changes that may be resolved by using the readout. *Accuracy* of calibration refers to the agreement or lack of agreement between the setting of frequency on the readout and the actual frequency produced. It has become accepted by industry that the term *standard-signal generator* means a generator of 1 percent or better frequency calibration accuracy. With the advent of electronic counters used as digital readout assemblies, it is now possible to achieve accuracies limited only by the number of digits provided and most importantly the accuracy of the time-base reference frequency provided. Many generators are also supplied with a frequency calibrator in the form of a crystal oscillator or oscillators to provide check points at discrete frequency intervals, usually 100 kHz or 1 or 5 MHz. With this, the generator can typically be set to 0.01 percent frequency accuracy at frequencies that are harmonically related to the calibration oscillator frequency.

Resettability is sometimes quoted by manufacturers and is a useful

and desirable specification, but again it should not be confused with accuracy. This specification indicates, usually in terms of frequency (hertz, kilohertz, etc.), how closely the output frequency will repeat when reset to a given frequency indicated by the readout assembly—no matter how much the readout may disagree with the actual output frequency.

Stability. The important characteristics relating to stability are residual FM, drift, and microphonism. In good signal generators, the peak deviation of the residual FM is less than 5 parts in 10^7 for frequencies less than 1 GHz, and from 3 to 15 kHz for higher frequencies. A drift of 10 ppm/10 min after a warm-up time of 30 min is typical in the lower frequency ranges. While changes in frequency caused by vibration and shock are extremely important, no adequate standards exist at present.

Harmonic Distortion. Total distortion that is less than 2 percent is usually acceptable in microwave generators, since it is easy to remove harmonics with low-pass filters.

Spurious Outputs (Nonharmonically Related Frequencies). This specification is more normally associated with synthesized signal generators where spurious outputs of at least 80 dB below the desired frequency are considered an adequate specification, but with the advent of the electronic counter readout in direct generators this specification has assumed increased importance. Receiver design, production testing, and maintenance are largely processes of determining and eliminating spurious responses, and the tests are accomplished by substituting a signal generator for the antenna signal. A signal generator with spurious outputs can therefore be an embarrassment to the user. Frequency-counter readouts can inject reference and time-base frequencies, or their harmonics, into the desired output, and thus signal generators having this readout means will include a spurious specification. An additional form of spurious output can be delivered by generators deriving their bands by division from a master tunable oscillator. In this case it is possible that subharmonics of the desired output will appear across the output terminals. These should also be specified separately, since by the very nature of their harmonic relationship they could cause difficulty in a given measurement application.

Output-level Characteristics. In addition to the information on standard-signal generators in Chap. 10, several things need to be said specifically about generators in the microwave range. For instance, electronic leveling of output as the tuning dial is moved is used more frequently in microwave generators than in those operating at lower frequency. Automatic leveling ranges from about ± 0.25 to about ± 1 dB across the dial.

Meter indication of output level into a matched load is required in standard-signal generators, as either voltage or, more often, power. Output levels ranging from greater than 0 dBm (0.223 V) to less than

−127 dBm (0.1 μV) are normally provided and very necessary in most applications. This is accomplished by providing at least 10 dB of continuous level control indicated by the meter and an attenuator of 10 db per step. The output level is then the sum of the meter reading and step-attenuator setting. Attenuator errors are normally specified in addition to meter accuracy and frequency response and therefore must be added to the metering error and frequency response error to obtain the total possible level error. Typical attenuator errors are 0.1 dB/10-dB step with a maximum accumulated error of 2 percent over a 120-dB range. At microwave frequencies these may be relaxed somewhat, but at frequencies below 1 GHz these should be considered maximum error specifications for a precision signal generator.

As previously mentioned, the output level is normally calibrated and specified into a matched load. For various reasons, 50 Ω is considered the standard source impedance for a signal generator, but in the lower frequency regions (< 500 MHz) both 60- and 75-Ω source impedances are available and others on special order. In any case a source vswr of less than 1.5:1 is to be expected.

Radio-frequency Interference or Leakage. Radio-frequency interference specifications are probably the least noticed ones on a signal-generator data sheet, and yet they are among the major specifications defining a standard-signal generator. Radio-frequency leakage from a signal generator can create practical difficulties in measuring receiver sensitivities, particularly if the receiver itself is not well shielded. Additionally, inaccuracies in attenuation at the microvolt level may result if strong radiation fields exist. Most rf engineers become acquainted with a similar problem very early in their careers when they attempt to put 150 db of attenuation in a coaxial cable. The usual result is that the signal level output of the cable shows no decrease in level for the last 30 dB or more of attenuation added. This result is obviously not due to the attenuator errors but to leakage from input to output connectors and from cables to the attenuator.

Precautions should be taken in the generator to ensure that radiation levels are sufficiently low to permit accurate measurements at the minimum output level specified for the signal generator. Coaxial cable, coaxial connectors, and rf shielding techniques used in the generator must be compatible in leakage with this minimum level.

Signal-generator manufacturers will indicate leakage performance in one of two ways, either by calling out an applicable specification or by stating that leakage levels permit measurements at the lowest output levels provided by the generator.

Modulation Characteristics. Space allows only a brief introduction to the modulation characteristics of signal generators. In general, a sig-

nal generator must provide some form of modulation, and it must be calibrated.

Four forms of modulation are typically provided on commercially available direct signal generators and synthesized-signal generators, but not all on the same generator or in the same frequency range. They are AM, video modulation (extended-frequency-range AM), FM, and pulse modulation.

Amplitude Modulation. Amplitude modulation is available on signal generators from 50 kHz to 40 GHz. A self-contained modulation oscillator should be provided, and metering to indicate depth of modulation is a necessity. Typically a 1,000-Hz modulation oscillator is provided, but some microwave generators provide a self-contained sine-wave modulation oscillator tunable from 20 Hz to 2 kHz. External AM must be available to allow other than sine-wave modulation, but when nonsinusoidal modulation is used, the modulation-depth meter reading is normally meaningless. At microwave frequencies (above 1 GHz) AM versatility is not needed and AM is usually only used in coherent detection measurements, but at frequencies below 1 GHz, AM is prevalent (Chap. 10) in communications systems.

When the AM bandwidth extends beyond 100 KHz and usually to many megahertz, the terminology is changed to video modulation. If this mode is provided, the generator versatility is greatly increased and the instrument becomes useful for measurements involving TV video modulation.

Pulse Modulation (PM). Pulse modulation is to the microwave frequency range what AM is to the lower frequency ranges. Historically microwave signal generators have included provisions for pulse modulation, but this has normally been accomplished by gating the oscillator on and off in some fashion with resulting problems of "moding" and high incidental FM. In recent years these limitations have been circumvented by the use of PIN modulators which absorb power on the transmission path. This technique allows the oscillator to operate continuously as in the lower frequency ranges and reduces reflections below levels that cause frequency pulling. For reasons of cost and versatility, high-quality pulse modulation for modern signal generators is usually provided by an accessory instrument.

Frequency Modulation (FM). Applications of FM within the range from 2 to 1,200 MHz are required of fixed and mobile communications, broadcasting, multiplexed telemetry systems, multichannel telephone links, and high-speed data transmission, including earth-space telemetry and meteorological data relay systems. FM has become in fact the most common method by which information is impressed on an RF carrier. It would not be economically feasible to provide the full modulation

requirements for all these applications in one signal generator, but many signal generators are available covering substantial portions of the FM applications. Because of this broad range of generators optimized for particular applications, only general specifications and characteristics will be discussed.

Peak deviation must be settable and accurately monitored over the full range provided—as a minimum, from 0 kHz to 0.2 percent of the carrier. The deviation, once set, should not require resetting with band change and should be monitored by a peak-reading meter to within at least 5 percent of full-scale accuracy. Modulation frequency response should be at least ± 0.5 dB and extend from below 20 Hz to above 150 kHz. As with AM, the higher-quality signal generators will usually provide a variable-FM oscillator. Typically a 1 percent distortion specification will satisfy most applications, but in addition, if stereo modulation is required, intermodulation distortion and group delay characteristics should be specified. Incidental AM occurring with the FM can be extremely troublesome, just as the incidental FM with AM, but fortunately is more easily removed with external limiters.

Frequency Stabilization in Signal Generators. In some applications the frequency stability of signal generators is not quite adequate. In this situation, an instrument (often called a *microwave-frequency stabilizer*) is available to synchronize the oscillator in a generator with some harmonic of an external frequency standard. It is only required that the original signal generator be tunable over a narrow range by means of a variable dc signal.

In a simple form the operation of a phase-lock frequency stabilizer is indicated in Fig. 15-36. A sample of the signal-generator rf is mixed with a comb of harmonics of the reference oscillator. The difference frequency which falls within the intermediate-frequency amplifier bandpass carries the frequency instabilities of the signal generator to be stabilized and the relatively smaller instabilities of the reference oscillator harmonic. The difference frequency is then compared in a phase detector against

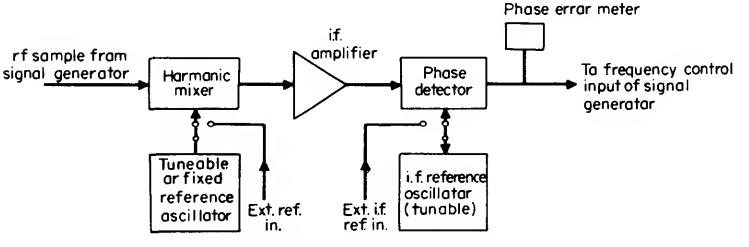


FIG 15-36 Simplified block diagram of a signal generator stabilizer.

either a crystal reference or a variable oscillator. The output voltage of the phase detector, which is proportional to the difference in phase of the two inputs and is the loop error signal, is then amplified and applied to the voltage-controlled reactance in the oscillator of the signal generator. Two special requirements, a sample of the rf output and an input for voltage tuning of the oscillator over a narrow range, must be provided.

The unique features of these instruments are the combination of continuous frequency coverage (in some cases discrete-step frequency coverage) and quartz-crystal spectral purity. In the phase-locked condition, the signal generator can assume both the long- and short-term stability of the reference oscillator. Provision is usually made for referencing the automatic-phase-control loop to any external quartz-crystal or atomic-referenced-crystal frequency standard. Thus, a signal generator when locked can have stabilities of 1×10^{-9} sec^{-1} to 1×10^{-9} day^{-1} or better depending on the reference employed. However, as with a synthesized signal generator, there are frequencies in the instrument not harmonically related to the desired frequency, and so spurious outputs are possible.

Synchronizers generally can lock oscillators over extremely wide frequency ranges in a single instrument and, in fact, in almost any signal generator that has provision for dc control of the oscillator. For example, synchronizers are commercially available which will lock over ranges of 50 kHz to 500 MHz and 1 to 40 GHz and thus span the ranges of many different and separate signal generators.

The resolution and settability of the system are still no better than the dial accuracy and resolution of the signal generator alone. Unless the user employs an electronic frequency counter to indicate the actual output frequency of the signal generator, he knows the absolute frequency with accuracy no better than that of the dial and is in fact worse off since the dc frequency control has usually pulled the oscillator off calibration. For this reason an electronic frequency counter is very commonly used as a dial for the signal generator when a synchronizer is used.

15-7 *Amplitude Modulators for Signal Generators*

Amplitude modulation is valuable not only for testing communication systems, but also for keeping the output level of a generator constant and for programming that output level. The signal level of any oscillator can be varied by varying the voltage on some electrode of the active oscillating device, but this approach to AM usually has at least two disadvantages. First, output amplitude rarely varies linearly with electrode voltage, and second, the frequency varies along with the amplitude.

A modulator that acts as a transmission device with variable attenu-

ation and allows the oscillator to operate steadily (continuous wave), under constant load, has many advantages.

Absorption Modulators. In 1962 the first absorption type of modulator based on the p-i-n diode was introduced as a product. This modulator is placed in the signal path and absorbs varying amounts of power depending on the modulation signal applied to the modulator. It is important that the modulator should always present a resistive and nearly constant impedance to the oscillator, since a reactive load, particularly one that varies with modulation signal, could also affect the frequency of the oscillator.

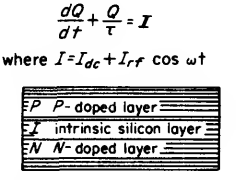


FIG 15-37 Cross section of PIN diode.

Let us first discuss the p-i-n (PIN) diode itself. The PIN diode is a planar, double-diffused junction semiconductor. The p and n impurities are diffused into opposite sides of a thin silicon wafer until only a thin layer of pure, intrinsic silicon remains between the two doped regions (see Fig. 15-37). This highly resistive silicon layer, typically 10 to 20 μm thick, gives the PIN diode its unique microwave properties. At low frequencies rectification takes place as it would with any p-n junction device. As the frequency is increased, however, the charge storage in the intrinsic region (or i layer) limits rectification. At ultrahigh and higher frequencies, rectification is practically nonexistent.

Diode conductance is proportional to stored charge q and inversely proportional to the square of the i-layer thickness. The stored charge is in turn related to diode current by

$$\frac{dq}{dt} + \frac{q}{\tau} = I \tag{15-7-1}$$

where

$$I = I_{dc} + I_{rf} \cos \omega t$$

and where τ is the carrier lifetime in the i layer. When the diode is driven by ac offset by an arbitrary bias level, its conductance g is

$$g = \frac{I_{dc}(\mu_n + \mu_p)}{d^2}\left[1 + \frac{I_{rf}/I_{dc}}{\sqrt{1 + \omega^2\tau^2}} \cos(\omega t + \phi)\right]$$

$$= g_{dc}\left[1 + \frac{I_{rf}/I_{dc}}{\sqrt{1 + \omega^2\tau^2}} \cos(\omega t + \phi)\right] \tag{15-7-2}$$

where ϕ = arctan $1/\omega\tau$
I_{dc} = dc component of drive current
I_{rf} = rf component of drive current $I_{rf} \cos \omega t$
τ = carrier lifetime in i layer, sec
μ_n = electron mobility in i layer, $cm^2/V\text{-sec}$
μ_p = hole mobility in i layer, $cm^2/V\text{-sec}$
d = i-layer thickness, cm
g_{dc} = dc conductance term

The time-varying term in Eq. (15-7-2) represents a nonlinear resistance to rf, but this term approaches zero for large $\omega\tau$. Thus long charge lifetime is desirable.

The lowest frequency for which the diode is usable is determined by the harmonic generation one is willing to accept. A typical diode might have a lifetime τ of about 100 nsec. With the diode biased so that its resistance is 100 Ω, the second harmonic would be about 30 dB down at 500 MHz. There are PIN diodes now available with lifetimes in excess of 1 msec, which makes them controllable microwave resistors at frequencies as low as 10 MHz.

An attenuating array can be formed by placing a number of PIN diodes across a transmission line. The diodes are spaced at quarter-wavelength intervals at the center frequency of the band of interest [36]. In a simplified analysis it can be assumed that the diodes are pure resistances. At zero bias the diode resistance is of the order of 5 to 10 kΩ. As bias current is increased, the diode resistance decreases. For an attenuating array they are never biased to less than about 2 Ω in order to maintain an adequate match. Input match can be further improved by tapering the diode impedances at the ends of the array. In Fig. 15-38 a tapered array is shown along with the input standing wave ratio as a function of frequency. The analysis used to derive these curves assumes that the array is infinitely long. The theory has been applied to arrays as short as seven diodes, however, with little difference between theory and actual results [37].

As was indicated, this analysis assumed the diodes to be pure resistances; this is a reasonably good assumption at the low end of the microwave region, but at higher frequencies the parasitic reactances due to the diode and its mounting structure must be taken into account.

Another class of modulators which has been extensively used in recent years makes use of the PIN diode as a reflective element. A schematic of this type of modulator is shown in Fig. 15-39. The PIN switches are simply two or four diodes mounted in shunt. The center conductor of the transmission line which interconnects the diodes is a very thin wire. This line looks inductive and at zero bias, in conjunction with the capacitance of the diode, forms a low-pass filter. The cutoff frequency of this

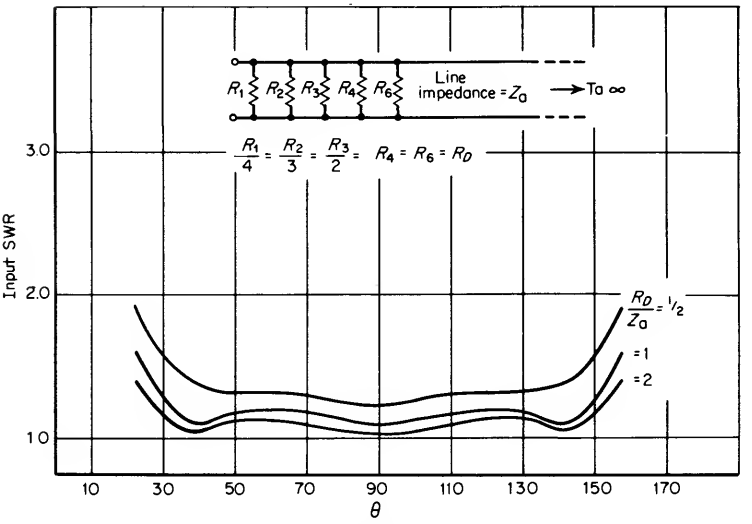


FIG 15-38 Attenuator array in transmission line.

filter can be made higher than 18 GHz, and this permits a well-matched structure up to 18 GHz. When the diodes are forward biased, their resistance can be decreased to almost a short circuit. At this point the PIN switch has a reflection coefficient almost equal to unity.

The operation of the modulator is as follows: Under zero-bias conditions power enters the input port and is split by the 3-dB coupler so that half the power goes through one switch and the remainder goes through the other switch and the signals recombine in phase at the output port. When the PIN switches are biased to a reflection coefficient of unity, the power reflects from the two switches and combines in phase at port 3 and is absorbed in the 50-Ω termination. Bandwidths of up to 4:1 and operation to 18 GHz have been achieved with this type of modulator.

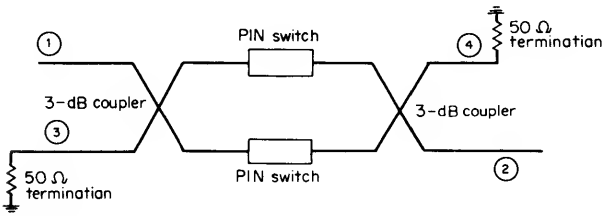


FIG 15-39 Reflective attenuator with the use of PIN diodes.

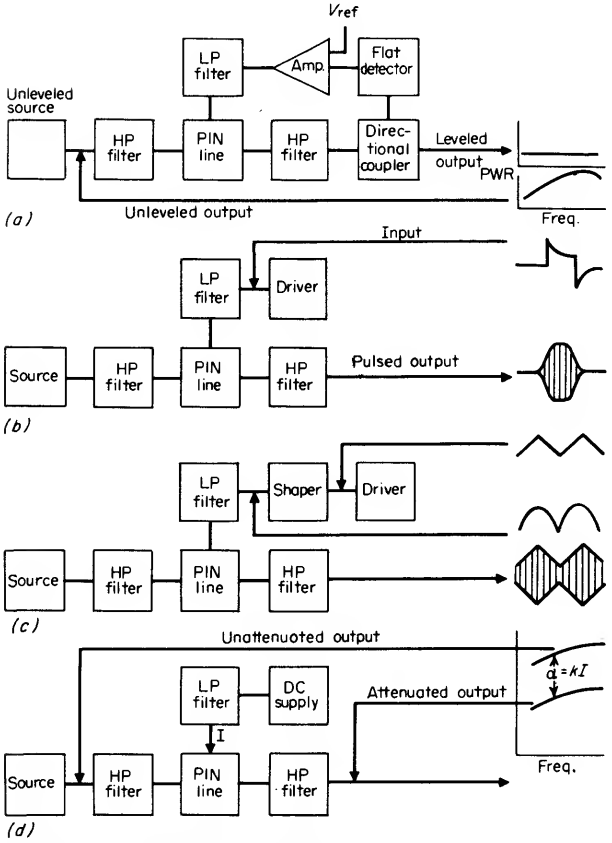


FIG 15-40 Applications of PIN lines: (*a*) leveling, (*b*) pulse modulation, (*c*) AM, (*d*) programmed attenuation.

Figure 15-40 shows some applications for PIN modulators (PIN "lines").

15-8 *Microwave Sweep Generators*

To test the many broadband microwave components and systems that now exist, it is convenient to have measuring instruments that automatically sweep through a range of frequencies. Swept signal generators, or sweepers, are therefore important, and the swept range should be at least an octave or the range over which a particular waveguide operates.

The backward wave oscillator is capable of being electrically tuned

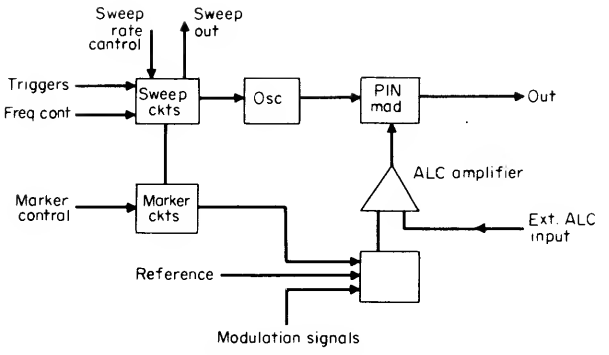


FIG 15-41 Block diagram for a typical microwave sweep oscillator.

over a considerable range. Its use is common in sweepers operating from 1 to 40 GHz, and it is even available for frequencies from 40 to 100 GHz. Voltage-tuned solid-state sources have also become available. These have been either varactor tuned or yig tuned, and for sweeper applications either method is acceptable. Solid-state sweepers are now available with yig- or varactor-tuned transistor oscillators up to 4 GHz and yig-tuned Gunn-effect devices from 4 to 12 GHz.

Many sweepers have plug-in capability so that the user can economically select his frequency band and plug in the proper oscillator assembly. In other instruments the various rf units can be selected by switches.

A simple block diagram of a sweeper is shown in Fig. 15-41, and Fig. 15-42 shows a typical measurement setup. The sweep output is a ramp voltage that varies linearly with oscillator frequency and is used to create a horizontal frequency axis on the cathode-ray oscilloscope. The vertical deflection in Fig. 15-42 is proportional to the gain magnitude of the device under test. As indicated in Fig. 15-41, the sweeper permits sweep

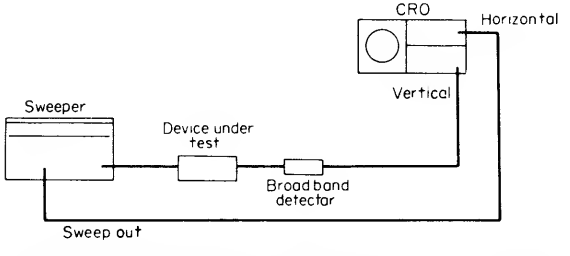


FIG 15-42 Typical setup for measurement of transmission coefficient to obtain a swept display.

rate, frequency limits, and output to be adjusted. Also, after the frequency has swept through certain increments, sharp marker pulses are injected into the PIN modulator to produce a visual frequency calibration on the oscilloscope.

Another essential aspect of a microwave sweep oscillator is the leveling circuitry. Leveling a sweep oscillator serves two purposes. Its primary function is to achieve flat output power as a function of frequency to facilitate broadband measurements. The other function that leveling serves is to improve source match, and this, as will be shown in a later chapter, is important too for accuracy in reflection coefficient and attenuation measurements.

To maintain a constant output level in the arrangement of Fig. 15-41, it would seem logical to connect the input of the automatic level control amplifier to the output terminal of the sweeper, so that feedback could force the output to be nearly equal to the reference. The trouble with this scheme is that the actual load is connected to the sweeper by means of a transmission line or system having arbitrary attenuation. It is better to derive a feedback signal at the point where leveled power is desired.

The schematic for a complete leveling loop is shown in Fig. 15-43. This circuit is not linear. The detector is generally operating at a level where it can be considered to have a square-law response so that $V_{\text{det}} = FkP_o$, where F is the coupling factor of the directional coupler and k is the detector sensitivity. The PIN modulator is approximately linear in dB of attenuation versus bias current, so that

$$P_o = P_s e^{-\alpha I} \tag{15-8-1}$$

where α denotes the sensitivity of the PIN line in nepers per milliampere. For small signals this characteristic can be approximated by

$$P_o = P_s(1 - \alpha I) \tag{15-8-2}$$

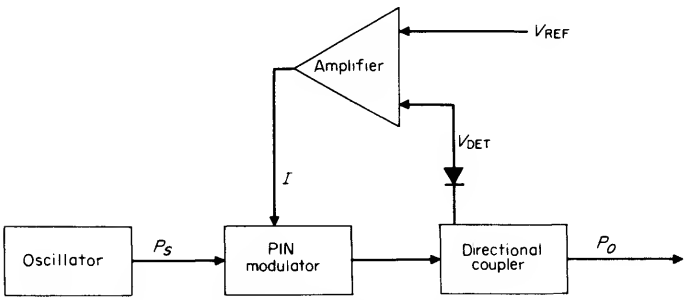


FIG 15-43 Block diagram for automatic level control.

Under these conditions solving for P_o yields

$$P_o = P_s \frac{1 + \alpha K V_{\text{ref}}}{1 + \alpha K F k P_s} \tag{15-8-3}$$

From this equation it can be seen that if the amplifier gain K is high,

$$P_o \approx \frac{V_{\text{ref}}}{Fk} \tag{15-8-4}$$

where α = PIN line sensitivity, nepers/mA
K = amplifier gain factor
F = coupling factor
k = detector sensitivity

A requirement, then, on loop gain is that it be high enough to make the above approximation. Another requirement is that bandwidth be sufficient to remove the most rapid variations that are likely to occur in P_s; a bandwidth of 50 kHz is generally sufficient. So far this discussion has assumed small signal conditions so that linearity could be assumed. Most leveling, however, must operate over a fairly large dynamic range. The automatic-level-control loop must operate over a dynamic range of more than 35 dB for most sweeper applications, and sufficient gain and bandwidth must be maintained over that range. Most sweepers have some provision for adjusting automatic-level-control loop gain so that the user can be assured that he is always operating at maximum possible gain without the occurrence of oscillations.

It is difficult to tune one microwave oscillator continuously over a range greater than about one octave. In sweeper applications, however, it is desirable to be able to sweep over a greater range, especially in applications below 1 GHz. A technique to get broader sweeps is to mix (heterodyne) the outputs of two oscillators, one or both of which is electrically tunable.

The block diagram for such a sweep oscillator unit is shown in Fig. 15-44. In this case one oscillator is tunable from 2.3 to 4.2 GHz and serves as the local oscillator for the mixer. The other oscillator is a fixed-frequency oscillator operating at low level compared with the level for the local oscillator, so as to minimize spurious mixing products. The mixer output power is proportional to the amplitude of the second oscillator; hence leveling and amplitude modulation can be accomplished by placing a modulator in the fixed-frequency signal path. The difference frequency at the output of the mixer now covers many octaves, but an inherent problem with this type of sweeper unit is the presence of spurious signals due to intermodulation in the mixer. In Fig. 15-45 a mode chart is shown that applies to the sweeper rf unit shown in Fig. 15-44.

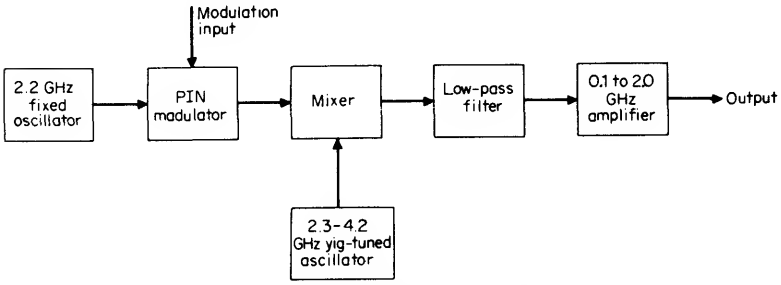
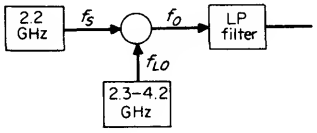


FIG 15-44 Block diagram for a Hewlett-Packard 8699B heterodyne sweeper plug-in.



$$f_O = \pm n f_{LO} \pm m f_S$$

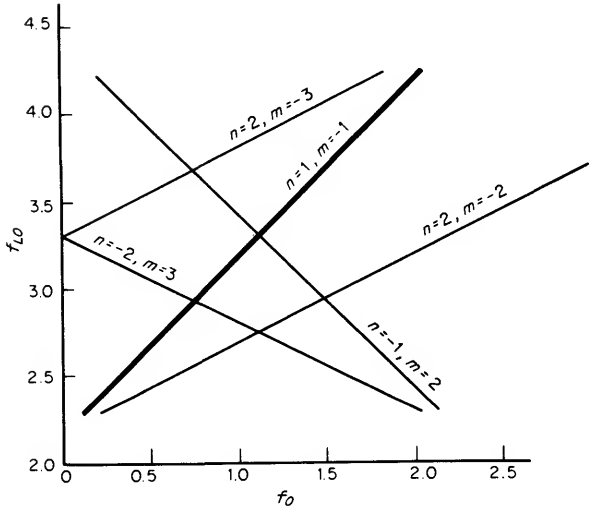


FIG 15-45 Mode chart for Hewlett-Packard 8699B: $n = 1$, $m = -1$ represents the desired mixing product; others represent unwanted spurious signals; in-band spurious signals up to the fifth order are shown; out-of-band mixing products are removed by the low-pass filter.

With proper mixer design and choice of signal levels, spurious signals can be kept 35 to 40 dB below the desired signal. At these levels, spurious signals very rarely interfere with measurement accuracies.

Another difficulty that arises in heterodyne sweepers is inaccuracy of frequency calibration. Since two large numbers are subtracted to get the output frequency, a small error in the frequency of either oscillator can result in a large error in output frequency. The frequency specification on the unit just described is ± 10 MHz, which is an error of 0.5 percent at 2 GHz but 10 percent at 100 MHz. At lower frequencies, a frequency discriminator can be used in a feedback loop to achieve frequency accuracies of 1 percent at the output frequency.

CITED REFERENCES

1. Mason, S. J.: Power Gain in Feedback Amplifiers, *IRE Trans. Circuit Theory*, *CT*-1, no. 2, pp. 20–25, 1954.
2. Ku, W. H.: Unilateral Gain and Stability Criterion of Active Two-ports in Terms of Scattering Parameters, *Proc. IEEE* (*Correspondence*), vol. 54, no. 11, pp. 1617–1618, 1966.
3. Cote, A. J., Jr.: Matrix Analysis of Oscillators and Transistor Applications, *IEEE Trans. CT*-5, no. 3, pp. 181–188, September, 1958.
4. Voltage Variable Capacitor Tuning: A Review, *Proc. IEEE*, vol. 56, no. 5, May, 1968.
5. Kompfner, R.: The Travelling Wave Tube as an Amplifier at Microwaves, *Proc. IRE*, February, 1947.
6. Gordon, J. P., H. J. Zeigler, and C. H. Townes: Molecular Microwave Oscillator and New Hyper-fine Structure in the Microwave Spectrum of NH_3, *Phys. Rev.*, vol. 95, no. 1, 1955.
7. Bloom, S., and K. K. N. Chang: Theory of Parametric Amplification Using Nonlinear Reactions, *RCA Rev.*, vol. 18, no. 4, December, 1957.
8. Esaki, Leo: New Phenomenon in Narrow Germanium p-n Junctions, *Phys. Rev.*, vol. 109, 1958.
9. Kurokawa, K.: Power Waves and the Scattering Matrix, *IEEE Trans.*, March, 1965.
10. Bodway, G.: Two-port Power Flow Analysis Using Generalized *S* Parameters, *Microwave J.*, May, 1967.
11. Bodway, G.: Circuit Design and Characterization of Transistors by Means of Three-port *S* Parameters, *Microwave J.*, May, 1968.
12. Network Analysis at Microwave Frequency, Hewlett-Packard Co., Appl. Note 92.
13. Wang, P. H. Y.: Design Considerations of a 0.1–2.0 GHz Wide-band Microwave I. C. Amplifiers, *Intern. Solid State Circuit Conf. Dig. Tech. Papers*, February, 1969.
14. Besser, L.: Combine *S* Parameters with Time Sharing, *Electron. Design*, vol. 16, Aug. 1, 1968.
15. Mason, S. J.: Feedback Theory—Some Properties of Signal Flow Graphs, *Proc. IRE*, vol. 41, September, 1953.
16. Page, C. H.: Frequency Conversion with Positive Nonlinear Resistors, Res. Paper 2664, *J. Res. Natl. Bur. Std.*, vol. 56, no. 4, April, 1956.

17. Manley, J. M., and H. E. Rowe: Some General Properties of Nonlinear Elements—Part 1, General Energy Relations, *Proc. IRE*, vol. 44, pp. 904–913, July, 1956.
18. Penfield, P., Jr., and R. P. Rafuse: "Varactor Applications," The M.I.T. Press, Cambridge, Mass., 1962.
19. Moll, J. L., and S. A. Hamilton: Physical Modeling of the Step Recovery Diode for Pulse and Harmonic Generation Circuits, *Proc. IEEE*, vol. 57, no. 7, pp. 1250–1259, July, 1969.
20. Burckhardt, C. B.: Analysis of Frequency Multiplier for Arbitrary Capacitance Variation and Drive Level, *BSTJ*, vol. 44, no. 4, pp. 675–692, April, 1965.
21. Tang, C. H.: An Exact Analysis of Varactor Frequency Multipliers, *IEEE Trans.*, vol. MTT-14, pp. 210–212, 1966.
22. Matthaei, G. L., L. Young, and E. M. T. Jones: "Microwave Filters, Impedance Matching Networks, and Coupling Structures." McGraw-Hill Book Company, New York, 1964.
23. Matthaei, G. L.: Tables of Chebyshev Impedance Transforming Networks of Low Pass Filter Form, *Proc. IEEE*, vol. 52, pp. 939–963, August, 1964.
24. Hamilton, S. A., and R. D. Hall: Shunt Mode Harmonic Generation Using Step Recovery Diodes, *Microwave J.*, pp. 69–79, April, 1967.
25. Hewlett-Packard Application Note 920, Harmonic Generation Using Step Recovery Diodes and SRD Modules, Hewlett-Packard Company, Palo Alto, Calif.
26. Special Issue on Semiconductor Bulk-effect and Transit Time Devices, *IEEE Trans. Electron Devices*, vol. ED-13, January, 1966.
    Second Special Issue on Semiconductor Bulk-effect and Transit Time Devices, *IEEE Trans. Electron Devices*, vol. ED-14, September, 1967.
27. Kennedy, W. K.: Power Generation in GaAs at Frequencies Far in Excess of the Intrinsic Gunn Frequency, *Proc. IEEE, Letters*, vol. 54, p. 710, April, 1966.
28. Copeland, J. A.: A New Mode of Operation for Bulk Negative Resistance Oscillators, *Proc. IEEE, Letters*, vol. 54, pp. 1479–1480, October, 1966.
29. Watson, H. A.: "Microwave Semiconductor Devices and Their Circuit Applications," McGraw-Hill Book Company, New York, 1969.
30. Hanson, Del: YIG Tuned Transferred Electron Oscillators Using Thin Film Microcircuits, *Dig. Tech. Papers Intern. Solid State Conf.*, February, 1969, p. 122.
31. Josenhans, J. G.: Noise Spectra of Read Diodes and Gunn Oscillators, *Proc. IEEE*, vol. 56, no. 4, pp. 762–763, April, 1968.
32. Read, W. T., Jr.: A Proposed High-frequency, Negative Resistance Diode, *Bell System Tech. J.*, vol. 37, pp. 401–446, March, 1958.
33. Young, Leo, ed.: "Advances in Microwaves," vol. 2, chap. 2, Academic Press, New York, 1966.
34. Barrett, M., and J. Viens: A Fully Integrated Varactor Tuned Avalanche Oscillator in Q Band, *Tech. Dig. Proc. Microelectronics Symp., Sept.* 10–11, 1969.
35. Leeson, D. B.: A Simple Model of Feedback Oscillator Noise Spectrum, *Proc. IEEE*, vol. 54, pp. 329–330, February, 1968.
36. Hunton, J. K., and A. G. Ryals: Microwave Variable Attenuators and Modulators Using PIN Diodes, *IRE Trans. Microwave Theory Tech.*, vol. 10, no. 4, p. 262, July, 1962.
37. Gray, D. A.: How to Design PIN-diode Control Devices, *Microwaves*, pp. 22–31, November, 1964.

CHAPTER SIXTEEN

MICROWAVE SIGNAL ANALYSIS

From notes by

Stephen F. Adam

Roderick Carlson

and Fred Pramann

Hewlett-Packard Company
Palo Alto, California

In the previous chapter, the generation of microwave signals for the purpose of measurement was studied. It follows naturally that the next step is to study the other measurement instruments and the techniques used for characterization of signals. Magnitude, frequency, harmonics, and modulation are some of the characteristics of microwave signals that one may need to analyze and describe quantitatively.

16-1 Power Measurement

At microwave frequencies power (rather than voltage or current) is the quantity commonly used to describe the magnitude of a signal. Sensors

responding to power are the more accurate at microwave frequencies. In addition, standing waves are usually present in a microwave transmission line of a magnitude to produce significant variations in voltage and current along the length of the line. The power available along the length of the transmission line is almost independent of distance, being diminished only by the small attenuation of the line. Power is a basic quantity and of primary interest to the microwave engineer.

A power meter can be a rather simple instrument, but the power sensing device associated with it deserves discussion, impedance matching must be given careful consideration, and some of the commercial power meters have clever circuitry worth studying.

Power Sensors. The power sensor is the key element of the instrumentation required to measure power at microwave frequencies. Most of these sensors provide a dissipative load for the energy and operate on the basis of heating. Usually we find a temperature rise in a thermocouple, a temperature-sensitive resistor, or a fluid. There are not many basic forms of sensors, but for each sensor there are several implementations to provide the desired power reading from the raw response of the sensor. The most important sensing devices will be studied below, and a few diagrams of instruments will be given.

Sensing by means of *thermocouples* has emerged as perhaps the best method for measuring microwave power. This sensor is composed of thin films of two thermoelectric metals that overlap slightly at the center of the device. The films are controlled in thickness to yield a resistance termination equal to the characteristic impedance of the transmission line to which it is connected. The sensor output is a dc voltage very nearly proportional to the microwave power dissipated. In the simplest embodiment, the thermocouple is connected to a dc amplifier with the right gain to cause the correct readout in power on a panel meter.

The thermocouple *mount* consists of the sensor elements, means for connecting to the microwave transmission line, and a cable carrying the thermocouple dc voltage to the associated instrument. The mount is carefully constructed so that the junctions between the thermoelectric materials and the lead conductors remain isothermal in spite of heat flow back and forth between sensor and the outside environment.

Usually two sensor elements are employed in a coaxial mount to make it possible to have a very wide bandwidth, such as 10 MHz to 18 GHz. Figure 16-1 shows the circuit configuration used. The low-frequency cutoff is determined by the blocking capacitor C_1 and the bypass capacitors C_2 and C_3. The high-frequency cutoff is limited by reflections within the system. Most coaxial power detectors utilizing other types of sensors, such as thermistors or barretters, also have two elements in the configuration of Fig. 16-1. This circuit has the advantage of providing

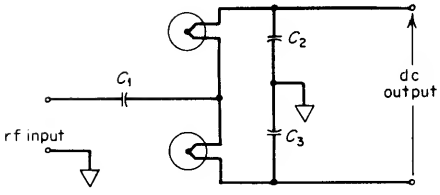


FIG 16-1 Circuit diagram of a thermocouple mount.

the two required contacts to the auxiliary power meter at rf ground, and the coaxial center conductor is loaded only by the desired sensor elements.

Figure 16-2 is a drawing of a thermoelectric subassembly containing two sensors bonded to a microcircuit on a sapphire substrate. The dissipation of microwave energy in the resistance of the bismuth and antimony films causes a temperature rise proportional to the power input and inversely proportional to the effective dissipation constant of the device. Figure 16-3 shows the trimetal system. The gold-bismuth (Au-Bi) and the antimony-gold (Sb-Au) junctions near the ends have a very low thermal resistance to the ambient temperature sink of the thermal system, and dc output drift is thus minimized. The bismuth-antimony (Bi-Sb) junction has a very *high* thermal resistance to the ambient heat sink. Therefore, the dissipation of a small amount of microwave power in the sensor resistance will result in a significant temperature rise in the area including the Bi-Sb junction. Since only the Bi-Sb junction is allowed to deviate from ambient, the output is

$$V_{dc} = K(S_{Sb} - S_{Bi})(T_j - T_{amb}) \qquad (16\text{-}1\text{-}1)$$

where K = ratio of thin-film thermoelectric power to bulk-metal thermoelectric power, $K \approx 0.8$
S_{Sb} = bulk thermoelectric coefficient of Sb, $+36\ \mu V/°C$
S_{Bi} = bulk thermoelectric coefficient of Bi, $-74\ \mu V/°C$
T_j = temperature of Bi-Sb junction
T_{amb} = temperature of the Au-Bi and Sb-Au junctions, which is also the ambient temperature of the mount.

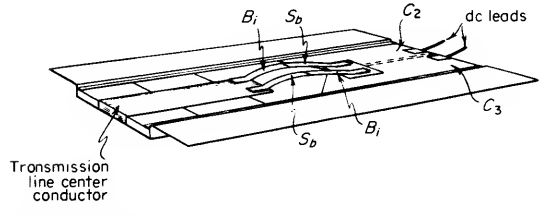


FIG 16-2 Two thermocouples mounted in the microwave supporting structure.

It can be seen from Eq. (16-1-1) that the dc voltage output of the device is proportional to the applied microwave power. The sensitivity of the sensor depends upon the difference in the thermoelectric coefficients of the pair of metals forming the hot junction of the sensor. When the

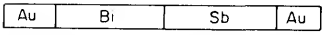


FIG 16-3 A microwave thermocouple.

metallic films are evaporated onto a substrate of poor thermal conductance to make the sensor sensitive, then the effective dissipation constant is a function primarily of the area of the resistor, since the only significant heat loss mechanism is conduction and convection through the air. An obvious disadvantage of a very sensitive device is that it is subject to burnout with a rather low level of power.

Thermocouple Power Meters. To complete the measurement system, a *power meter* is added to the thermocouple detector. This instrument typically contains the amplifiers, control circuitry, and calibration means required to produce an accurate indication of input power on a panel meter. In addition some instruments provide outputs for an electric recorder or DVM.

The dc amplifier, to amplify the output of the thermocouple sensor, requires special consideration. For one thing, the thermocouple voltage is very low; approximately 10 mV corresponds to full scale on the *highest* power range. Of course, one should like to achieve the lowest possible range, and the maximum practical sensitivity is limited by either thermal agitation noise in the thermocouple resistance or drift in the dc amplifier.

Currently, the dc amplifier with the lowest drift for use with source impedances of 200 Ω is one that uses a mechanical chopper, or modulator. The input circuit in Fig. 16-4 shows such a chopper feeding a transformer. Current from the sensor flows first down through the top primary winding and then up through the bottom winding, which causes almost a triangular wave of flux in the core and a square wave of output voltage. A high step-up ratio is used to keep the noise in the first stage of ac amplification from being more significant than the noise in the source resistance, and the transformer is very well shielded. The FET used in the first amplifier stage provides high input impedance and is chosen for low noise.

The shielding requirement is reduced, however, by operating the chopping switch at 110 Hz, which, together with synchronous demodulation, avoids intermodulation with harmonics of either 50- or 60-Hz line frequencies. The arrangement in Fig. 16-4 is presently capable of a full-

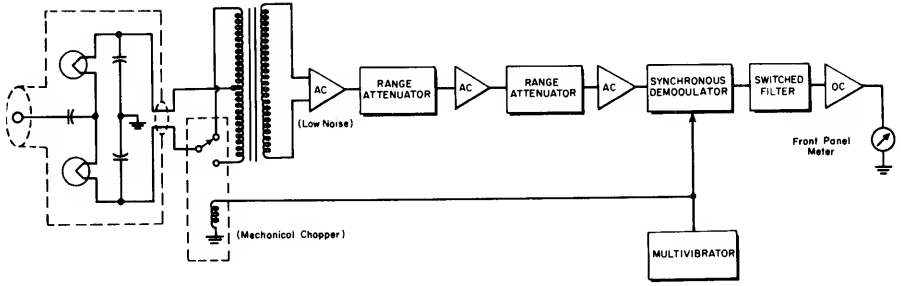


FIG 16-4 Block diagram of a thermocouple power meter.

scale setting of 300×10^{-9} V without drift being the main limitation. Transistor choppers will not meet the drift requirement, and photoconductive choppers have too much thermal noise, although they are useful for high-impedance sources.

If the amplifier is adequate, sensitivity is limited by thermal noise generated in the 200 Ω of resistance in the thermocouples; this input noise at room temperature is $(1.8 \times 10^{-9}$ rms V$)/B^{1/2}$, where B is the noise bandwidth of the system in hertz. In a modulation system such as the one being described, effective bandwidth is determined by the low-pass filter after the synchronous demodulator. The noise level on the lower ranges can readily be decreased by band limiting in this filter, but only at a sacrifice of indicating or recording speed. Unfortunately, the voltage of white noise varies only as the square root of bandwidth.

A method of easily calibrating the power meter is desirable, for no two thermocouple mounts have exactly the same power conversion sensitivity. In fact, the sensitivity of any particular mount varies with change of ambient temperature. Calibration is easy, however, because the thermocouples respond to audio frequencies as well as microwave frequencies. The power meter can be obtained with a self-contained audio generator of good amplitude stability. A "calibrate" switch is used to apply a known audio power to the two couples in series while microwave input is zero, without interfering with the performance of the amplifying system. The gain of some part of the system is adjusted to give the correct reading on the output meter while the calibration signal is applied.

Thermistor Power Meters. Thin-film techniques had to be developed to construct the tiny Bi-Sb couples described above. Previously, *thermistors* were predominantly used as sensors; thermistor power meters are still competitive with couple meters and even have some advantages.

Basically, a thermistor is a resistor with a large negative temperature coefficient. Usually the component is composed of a sintered mixture of various metal oxides. Figure 16-5 shows a sketch of a typical thermistor

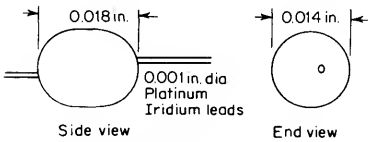


FIG 16-5 Microwave bead thermistor.

bead for use in a mount for the measurement of microwave power; its small size makes it difficult to produce with uniformity. Heating the bead from room temperature to about 140°C makes its resistance fall from about 1,500 to 100 Ω. It is operated in practice at 100 Ω.

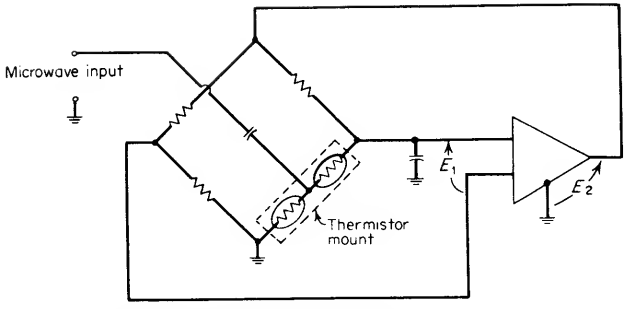


FIG 16-6 Self-balancing thermistor bridge.

Two thermistor beads are generally used in series in a microwave mount. This gives a terminal at the junction between them where a coaxial microwave connection can be conveniently made to feed the two in parallel, as in Fig. 16-6. Several sophisticated circuits have been developed to obtain accuracy and immunity from changes in ambient temperature, and one such circuit will be described later, but an elementary scheme is shown in Fig. 16-6. Observe that the thermistor bridge is excited by an amplifier that is fed by the bridge output E_1. Figure 16-7 shows E_1 versus the excitation voltage E_2 in a qualitative manner. Assume that the gain K of the dc amplifier in Fig. 16-6 is high enough to

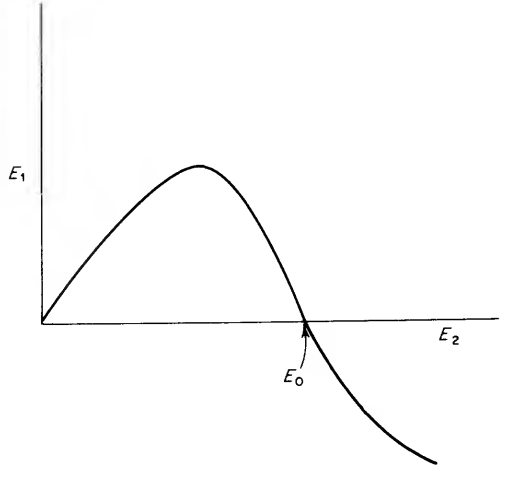


FIG 16-7 Approximate curve of output versus input voltage in a thermistor bridge.

drive E_2 to the vicinity of E_o, the excitation required for null. Then

$$E_1 = -\beta(E_2 - E_o) \tag{16-1-2}$$

where $-\beta$ is the slope of the bridge characteristic near null. But $E_2 = KE_1$, and therefore

$$E_1 = -\beta K E_1 + \beta E_o = \frac{\beta E_o}{1 + \beta K} \approx \frac{E_o}{K} \tag{16-1-3}$$

when $\beta K \gg 1$. The use of a high-gain amplifier will keep the bridge essentially at balance. An ac amplifier can be used instead of dc.

When microwave power is fed to the thermistors, the circuit will keep the bridge balanced by decreasing the power from the amplifier to the thermistors an amount equal to the microwave power, since the summation of all electrical heating of the thermistors must not change if the resistance of the thermistors is to remain constant. Therefore, the square of the change in E_2 is a measure of input power. Several methods have been devised to measure ΔE_2.

The main trouble with the simple circuit of Fig. 16-6 is that it is sensitive to changes in ambient temperature. To a good approximation, the power required to bias the thermistor to the proper operating temperature is

$$P_b = K_d(T_o - T_a) \tag{16-1-4}$$

where T_o = operating temperature, °C
T_a = ambient temperature, °C
K_d = dissipation constant of the pair of beads, 0.27 mW/°C in practice

The change of P_b with respect to T_a is

$$\frac{dP_b}{dT_a} = -K_d \tag{16-1-5}$$

Therefore in the output reading of the power meter, drifts are equivalent to 0.27 mW for each degree Celsius change in ambient, since the thermistor cannot distinguish between microwave power and power loss to the air. This is a severe drift in a power meter in which the *highest* power measured is 10 mW.

The tendency to drift can be reduced by two orders of magnitude by placing an additional set of thermistors in the same environment as the active thermistors and using them to compensate for ambient temperature changes. A great deal of art is required, since the two sets of thermistors must have nearly identical dissipation constants and operating environments. A good example of a power meter with this kind of

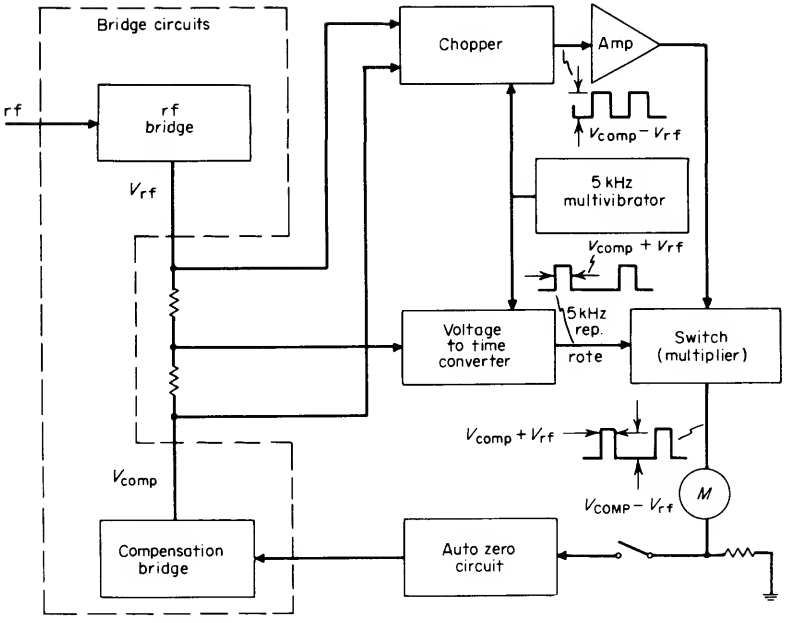


FIG 16-8 Block diagram of a thermistor power meter.

compensation will be discussed in connection with Fig. 16-8 (Hewlett-Packard model 432A).

Two matched sets of thermistor beads are located in the mount or sensing head, and two complete dc self-balancing bridge circuits similar to those in Fig. 16-6 are used, though the circuit details are not shown. Both the sum and difference of the two bridge output voltages are made available by a resistance network.

Both bridges produce self-balancing output voltages proportional to the square root of the power introduced into their respective thermistors. The power into one bridge (compensating bridge) is P_C, or compensating power, and the power into the other bridge (rf bridge) is P_{rf} plus P_M, the power to be measured. Therefore, the difference between the outputs of the two bridges can be used to measure the true rf power input if the two bridges are identical, or

$$P_C = P_{rf} + P_M \tag{16-1-6}$$

$$P_M = P_C - P_{rf} = K(V_C^2 - V_{rf}^2) = K(V_C + V_{rf})(V_C - V_{rf}) \tag{16-1-7}$$

where the subscripts on the voltages correspond to the bridges that produce the voltages and K is a constant.

The purpose of the remainder of the circuitry is to solve Eq. (16-1-7). The difference voltage, which can be extremely small, is amplified by a chopper amplifier driven by a 5-kHz multivibrator. The output of this amplifier is a square wave, as shown, since it is not demodulated. The sum voltage is used to control a voltage–to–pulse-width converter that is operated in synchronism with the difference chopper. This pulse width determines how long the gate is opened in the switching multiplier. Therefore, the output of the multiplier is a pulse train with pulse width proportional to $V_C + V_{rf}$ and pulse height proportional to $V_C - V_{rf}$. The area under the pulses is obviously proportional to the product, which is proportional to measured power. Only an averaging meter or circuit is needed to indicate this power.

Other Sensors. Historically the earliest way to measure microwave power was to use a solid-state rectifying *diode* to measure the voltage across a known resistance. At low signal level, diodes are nonlinear rectifiers, the actual voltage-versus-current curves depending upon the choice of diode: point-contact, ordinary junction, or hot-carrier junction. In any case, the proper choice of operating level and load produces a dc output current proportional to the square of ac input voltage from a low-impedance source, over a fairly wide dynamic range. This characteristic is convenient in the measurement of power with simplicity. However, the variability in rectification characteristic and frequency range limited by junction capacitance and carrier mobility in the diode does limit the adaptability of this type of power meter. Also, diodes are subject to mechanical damage.

The *barretter*, a tiny metal resistive element with a positive temperature coefficient, has been used extensively in the past to measure power. Its most common physical form is a short platinum wire with a diameter of only 30 to 60 μin., and it is used in a power meter with some sort of resistance bridge. At least one such power meter is a self-balancing bridge that accepts either a barretter or thermistor mount by having a switch that reverses the phase of the balancing amplifier to take care of a sensing element with either a positive or negative temperature coefficient.

In order to get sufficient sensitivity in a barretter bridge, it is necessary to operate the sensor at a bias power that is about half the power level sufficient to burn it out.

16-2 Measurement of Power Exceeding 100 mW

The measurement of power in excess of the range of a power meter is often achieved by using an attenuator or directional coupler to reduce the level of the signal. The attenuator usually provides a significant decrease in accuracy not only because of the uncertainty of the attenu-

ation, but also because of the impedance mismatch ambiguities (to be studied later) at both the input and output ports of the attenuating device.

The directional coupler has the disadvantage of requiring calibration of the coupling factor for accurate measurements. The coupling factor varies with frequency. Directional couplers will be studied in Chap. 17.

For the *direct* measurement of power levels above 100 mW, calorimeters are most often used. That is, the power being measured is dissipated in a thermal device, and the resulting temperature rise is observed. Classical calorimetry consists of observing the temperature rise in a well-insulated thermal mass of known thermal capacity after power is absorbed in that mass for a known time. This is more directly a measurement of energy, and a different approach is commonly used in the measurement of electrical power. Usually, the unknown power is made to cause a steady temperature rise at some point in a thermal network. This may be a temperature rise of a fluid of known physical characteristics and flow rate, a fluid carrying the heat away from the dissipative element. Or the power may establish a temperature gradient in a physical structure of known, fixed characteristics.

16-3 Pulsed-power Measurements

In the material above, it has been assumed that either the high-frequency power being measured was unmodulated or the average power in a modulated signal was desired. Pulsed rf power introduces some special problems in measurement, and the measurement of peak pulse power has been very important since the early development of pulse radar.

Peak Power Calibrators. Probably the most direct approach is to attempt to use a power sensor with a response so fast that its output faithfully follows the pulse modulation. This approach rules out thermistors, barretters, and thermocouple sensors at present because of their long thermal time constants. Semiconductor crystal diodes can be made to follow rf pulses with widths as narrow as 0.25 μsec with repetition frequencies as high as 1.5 MHz. If their variation in curvature of response characteristic and in sensitivity can be tolerated or taken into account, as well as their rf limitation, they are useful in the present application.

Figure 16-9 is a block diagram of such a "calibrator." Essential to most peak power measurements is a power divider (usually a precision directional coupler with a known coupling factor) since the peak power in pulses is usually much higher than the maximum ratings of the sensors described above. The instrument on which Fig. 16-9 is based reads directly to 200 mW, and external power dividers and terminations can be used to extend usage to much higher levels.

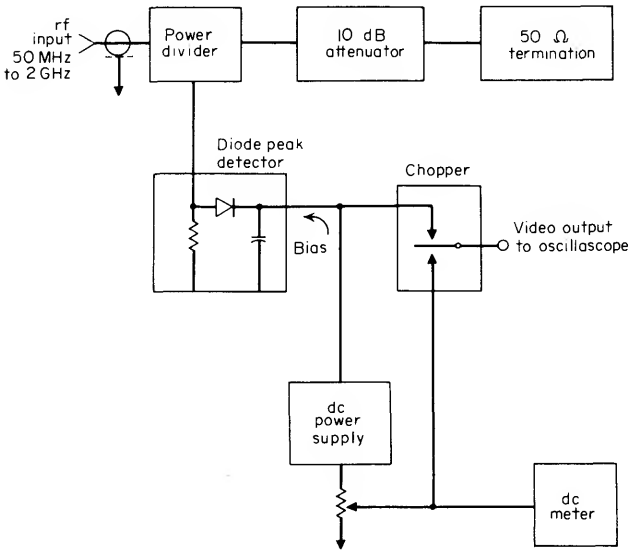


FIG 16-9 Block diagram of a peak-power calibrator that uses a crystal detector to measure actual peaks in pulse modulation.

A small fraction of the input power in Fig. 16-9 is applied to a diode peak detector, forward biased to stabilize the rectification characteristic and make the response to power almost linear. A chopper (sampling switch) alternately applies the diode output and the output of a calibrated dc power supply to an oscilloscope. With proper synchronizing, the scope shows two traces, one for each signal. If the positions of the two traces overlap when no power is being measured, and then the dc reference trace is adjusted to touch the peak of the pulse trace, the dc meter can be used to read out the peak pulse power.

Of course, this assumes that the instrument has been calibrated. To make the calibration requires only that an accurate continuous-wave power meter be substituted for the 50-Ω termination and a calibrated continuous-wave signal generator be connected to the input terminals. The internal 10-dB attenuator permits a wider selection of commercial power meters. The upper frequency range of peak power meters of this general type is limited to about 2,000 MHz.

Pulse-power Measurement with Slow Sensors. There are two general methods for using slow sensors to measure the peak power in rf pulses. One method is to make a waveform compensation for the distortion produced by the time constant of the sensor. There are commercial instruments with barretters that have been designed around this principle. The barretter is so slow (time constant of from 100 to 200 μsec) that its

output is virtually the integral of a square rf pulse. That is, the output is a straight line with a slope proportional to the peak pulse power. To reconstruct the pulse from its integrated form requires an active or passive circuit that differentiates the output of the barretter bridge. With adequate calibration techniques, this scheme is capable of an accuracy of better than 0.8 dB in power, and while this may seem like considerable error, the rf range extends throughout the common waveguide bands. The principal errors are associated with barretter tolerance and linearity, mount efficiency, and differentiating amplifier accuracy.

The other way to use slow sensors for peak pulse measurement is to measure average power in the repetitive pulses and use some auxiliary method to measure the duty cycle of the pulses, which is the product of pulse width and repetition frequency. Attenuating directional couplers are used to reduce the measured power to a convenient level. The peak power out of the coupler into the instrument setup is

$$P_p = \frac{\text{average power}}{\text{duty cycle}} \tag{16-3-1}$$

Pulse width in this method is most easily measured with a crystal detector operating in a condition to give it the fastest response to an rf pulse. The detector output is fed to a calibrated oscilloscope, where width and repetition frequency are observed.

Detailed information on the instruments and systems described tutorially here can be obtained from instrument manufacturers prominent in the microwave field.

16-4 Mismatch Considerations

The effects of mismatch at load and source must be considered in order to make accurate measurements and in order to state quantitatively the magnitude of the mismatch uncertainty. (The term *uncertainty* is used here instead of the term *accuracy*. One may speak of a large or small uncertainty but not a large or small accuracy.) The error due to mismatch is often the largest error in a power measurement.

The power delivered to a load by a source is determined by the general equation

$$P_D = \frac{(1 - |\Gamma_G|^2)(1 - |\Gamma_L|^2)}{|1 - \Gamma_G\Gamma_L|^2} P_A \tag{16-4-1}$$

where Γ_G = reflection coefficient of generator
Γ_L = reflection coefficient of load
P_A = power available from source
P_D = net power delivered to load

The quantities Γ_L and Γ_G are complex. The usual representation in polar form is $\Gamma = \rho e^{-j\theta}$.

Conjugate Match. The condition for conjugate match is that the load reflection coefficient be the complex conjugate of the source reflection coefficient. If $\Gamma_G = \rho_G e^{-j\theta}$, then Γ_L must equal $\rho_G e^{+j\theta}$ to be equal to the complex conjugate of Γ_G.

Substituting in Eq. (16-4-1) yields $P_D = P_A$ when a conjugate match is achieved.

Z_0 Match. A second frequently encountered condition of match is that of power delivered to a nonreflecting load (impedance equal to the characteristic impedance of the line, Z_0). This requires that $\Gamma_L = 0$. Thus Eq. (16-4-1) becomes $P_D = (1 - |\Gamma_G|^2)P_A \equiv P_o$, the power delivered to a nonreflecting load.

Mismatch Errors. If neither of the foregoing conditions is satisfied and in addition $\Gamma_G \neq 0$, then Eq. (16-4-1) is not readily solved, because usually arg Γ_G is not known and this may also be true of $|\Gamma_G|$. The Γ_L can be measured with a network analyzer. When arg $\Gamma_G\Gamma_L$ is not known, the limits of the quantity $|1 - \Gamma_G\Gamma_L|^2$ should be noted:

$$(1 - |\Gamma_G| \cdot |\Gamma_L|)^2 \leq |1 - \Gamma_G\Gamma_L|^2 \leq (1 + |\Gamma_G| \cdot |\Gamma_L|)^2 \qquad (16\text{-}4\text{-}2)$$

If $|\Gamma_G| = 0.2$ (swr = 1.5), $|\Gamma_L| = 0.13$ (swr = 1.3), and arg $\Gamma_G\Gamma_L$ is unknown, then the limit on $|1 - \Gamma_G\Gamma_L|^2$ is $(1 \pm 0.26)^2$, approximately a ± 5.3 percent error.

Generator Reflection Coefficient. It is apparent that the mismatch error will be eliminated if $\Gamma_G = 0$ and will be small if Γ_G is small. Attaching an isolator to the generator will give a value of Γ_G nearly equal to the output Γ of the isolator. This has a second advantage of isolating the generator from changes in load, which may "pull" the frequency or signal amplitude.

16-5 *Application Considerations*

Let us now consider the problem of measuring power under the two most frequent matching situations. A system would be designed for conjugate match when it is desired to get maximum power delivered to a load from a source of limited output. A good example would be maximizing the power delivered from a radar transmitter to the antenna. It is likely that both the transmitter and antenna have tuning elements to match them to the line so that excessive standing-wave ratios will not exist on the transmission line.

Most other equipment is designed to be matched to the characteristic impedance of the line. This facilitates standardization, measuring techniques, and compatibility. For instance, it is very desirable to design a

signal generator so that it presents a source impedance close to Z_0. The most meaningful measurements are made when devices with input impedances approximating Z_0 are attached. It is logical then to base the calibration of the generator on the power it will deliver to a nonreflecting (Z_0) load.

A power meter can only measure the power dissipated in the sensing element, but this is not all the power entering the sensing mount. Some power is absorbed by conductor surfaces, supportive dielectric materials, and blocking or bypass capacitors. The mount efficiency has been defined as the ratio of rf power absorbed by sensor elements to rf power dissipated within the mount as a whole. However, the power absorbed by the elements is not nearly as well known as the substitution power at dc or low frequency that is required to balance a bridge or bring a compensating bridge to the same operating point that a measuring bridge has. With thermocouples, the low-frequency power produces the same dc output as the microwave power. Therefore, the useful term *effective efficiency* has been defined as the ratio of substituted dc or low-frequency power to total rf power absorbed by the mount unit. It is this substituted power P_S which is the signal operated upon by the electronics in the power meter. The effective efficiency is

$$\eta e = \frac{P_S}{P_D} \tag{16-5-1}$$

There will be some error in the electronics of the power meter; therefore the power indicated P_{ind} will not be a perfect representation of P_D from the elements because of the instrumentation error e. Therefore,

$$P_{\text{ind}} = (1 - e)P_S \tag{16-5-2}$$

When making a measurement of the maximum power available from a source, it is necessary to match the mount and source by using a tuner. Define the tuner power transmission characteristic K_T to be the ratio of the power that the tuner can deliver to a load (connected to it) to the net power delivered to the input port of the tuner. In a good tuner, $K_T \approx 0.97$, which is difficult to measure and varies with probe penetration.

Available-power Measurement with a Conjugate Match. By combining the terms of this section with Eq. (16-4-1) to find the relationship of P_A and P_{ind} when the tuner is adjusted for a conjugate match,

$$P_{\text{ind}} = (1 - e)K_T\eta e P_A \tag{16-5-3}$$

Comparison Measurement with a Conjugate Match. If the source just measured is in turn connected to the desired load and tuned for a con-

jugate match,

$$P_{\text{load}} = K_{T2}P_A = \frac{K_{T2}P_{\text{ind}}}{K_{T1}\eta e(1 - \eta e)} \tag{16-5-4}$$

It is to be noted that the power transmission characteristic of the tuner must be evaluated for both settings.

Z_0 Power Measurement. If the mount used to measure the power has a reflection coefficient Γ_M that is not equal to zero, by substituting the terms of this section in Eq. (16-5-3), we have

$$P_{\text{ind}} = (1 - e)\eta e \frac{(1 - |\Gamma_G|^2)(1 - |\Gamma_M|^2)}{|1 - \Gamma_G\Gamma_M|^2} P_A \tag{16-5-5}$$

However, we want P_{ind} in terms of $P_o = (1 - |\Gamma_G|^2)P_A$.

Also it is convenient to use calibration factor K_b instead of ηe.

Calibration factor is defined as the ratio of substituted dc or low-frequency power dissipated in the elements to the rf power incident upon the mount. It differs from effective efficiency by the mismatch loss at the input coupling

$$K_b = (1 - |\Gamma_M|^2)\eta e \tag{16-5-6}$$

Rewritten,

$$P_{\text{ind}} = \frac{(1 - e)K_bP_o}{|1 - \Gamma_G\Gamma_M|^2} \tag{16-5-7}$$

Z_0 Comparison Power Measurement. If the source just measured is connected to a load, the power delivered to the load, P_L, is

$$P_L = \frac{(1 - |\Gamma_L|^2)|1 - \Gamma_G\Gamma_M|^2}{K_b(1 - e)|1 - \Gamma_G\Gamma_L|^2} P_{\text{ind}} \tag{16-5-8}$$

In both measurements, the great importance of small $\Gamma_G\Gamma_M$ and Γ_L is readily seen. The fact that arg $\Gamma_G\Gamma_M$ and arg $\Gamma_G\Gamma_L$ are not known is usually the cause of the largest uncertainty in the measurement.

The use of a tuner and the application of effective efficiency should be associated with a conjugate measurement, while the absence of a tuner and the application of calibration factor should be associated with Z_0 power measurements.

Errors in Sensor Mounts. These errors, or uncertainties, are usually specified separately from instrument errors, which vary from ± 0.5 to ± 3.0 percent. The *effective* efficiency of a mount, as defined above, is frequently determined by sending the mount to the National Bureau of Standards for calibration or by comparing it with a mount which is

traceable to NBS. The uncertainty of NBS calibration is from 0.5 to 2 percent, depending on the specific frequency and whether it is a waveguide or a coaxial mount. If a mount is compared with an NBS-traceable mount, the uncertainty is the sum of the uncertainty of the transfer measurement and the uncertainty of the NBS-traceable mount. By the definition, effective efficiency is independent of the mount reflection coefficient. Also, it should be independent of the power level. It tends to be stable with time and environment, but should be checked periodically.

Most commercial mounts are individually calibrated for effective efficiency by the manufacturer at six or seven microwave frequencies. The worst uncertainties are ± 2 to ± 5 percent, depending on frequency and calibration technique. By taking the square root of the sum of squares of the individual uncertainties making up the worst example of uncertainty, we obtain probable uncertainties of ± 0.5 to ± 2.5 percent. The effective efficiency is a function of frequency. A coaxial mount is typically 0.99 at 0.1 GHz and falls off to about 0.94 at 18 GHz. For waveguide mounts, 0.98 is typical with a small frequency dependence. Most commercial power-measuring equipment has means for dialing in the effective efficiency or calibration factor appropriate to the frequency of the source, and this adjusts the gain of the power meter so that correction is applied.

As with effective efficiency, calibration factor is determined by traceability to standards and NBS. The uncertainty of NBS calibrations ranges from 0.5 to 2 percent. Most commercial mounts are also individually calibrated for calibration factor. The range of uncertainties is about the same as for effective efficiency. To interpolate to get calibration factor at some frequency between two frequencies at which it is known, it is most accurate to interpolate the effective efficiency, measure the reflection coefficient $|\Gamma_m|$ accurately at that frequency, and determine the calibration factor from Eq. (16-5-6).

Dual-element error is found in mounts that have two sensor elements in parallel to microwave power, but are in series for the substitution power. Errors of a few percent have been observed in thermistor mounts. However, the technique of thermal matching used by some manufacturers in temperature-compensated mounts limits the error to about 0.1 percent in these mounts. The dual-element error is significant only on the highest range or two of a power-measuring system.

16-6 Microwave Frequency Meters or Wavemeters

The measurement of frequency by heterodyning, counting, and comparing with frequency standards has been treated in Chap. 6. The greatest accuracy of frequency measurement is obtained by the electronic techniques described in that chapter, because comparisons with standards

can be easily made to, say, 1 part in 10^{10}, but there is a large market for simple inexpensive wavemeters that operate on the basis of electrical resonance in mechanical structures. These structures are tunable.

While these tunable resonant wavemeters (or frequency meters) are basically simple, they are fabricated with a precision that keeps errors as low as a few parts in 10^5 in the best instruments.

When a resonant circuit is coupled to a source of microwave power, it of course absorbs the greatest power at the frequency of its resonance. As the frequency of the source is made to differ from the resonant frequency of the tunable device, the absorbed power decreases at a rate determined by the Q of the structure.

To obtain both high Q and precision of tuning at microwave frequencies, various versions of open-circuited or short-circuited transmission lines are used instead of lumped reactances. Any kind of transmission line can be used, including waveguide, but we shall call them all *cavity resonators* in this chapter. The input impedance of a short-circuited, low-loss transmission line is

$$Z_i = jZ_0 \tan \beta l \tag{16-6-1}$$

where Z_0 = characteristic line impedance
l = length of line
β = phase constant = $2\pi/\lambda$
λ = wavelength of signal

The above equation is an approximation for lines that are long in comparison with their other dimensions. The impedance Z_i approaches zero for even numbers of quarter wavelengths, and it approaches ∞ for odd numbers. Half-wave shorted lines are frequently used, and Fig. 16-10 shows a coaxial example with a short circuit adjustable by a plunger. The resonant frequency of such a half-wave cavity is

$$f_r = v\sqrt{\left(\frac{n}{2l}\right)^2 + \left(\frac{l}{\lambda_c}\right)^2} \tag{16-6-2}$$

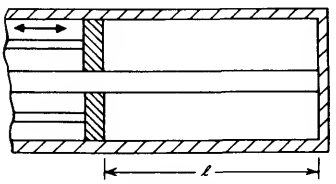


FIG 16-10 Coaxial half-wavelength cavity.

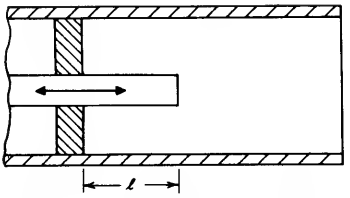


FIG 16-11 Coaxial quarter-wave-length cavity.

where v = velocity of propagation in the line
n = an integer
λ_c = cutoff wavelength of the line
which approaches ∞ for the principal transverse electromagnetic mode (TEM).

If one end of the cavity is left open as in Fig. 16-11, a quarter-wave cavity results, one that is resonant at odd multiples of quarter-wavelengths. In this construction, it seems more convenient to make the center conductor of the coaxial system movable than to slide the shorting element in the cylinder.

While a half-wavelength coaxial cavity can cover only a 2:1 frequency range without having any spurious resonances in its bandwidth, a quarter-wavelength coaxial cavity is capable of covering a 3:1 bandwidth, because its next higher resonance will be the $3\lambda/4$ one. That is, resonance occurs on a quarter-wavelength cavity at

$$l = \frac{\lambda_0}{4}(2n - l) \tag{16-6-3}$$

where λ_0 is the basic or longest wavelength for resonance.

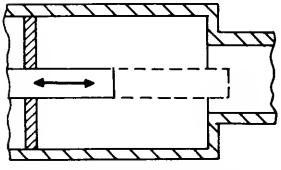


FIG 16-12 Capacitively loaded coaxial quarter-wavelength cavity.

The bandwidth free of spurious resonances can be increased in a quarter-wavelength cavity by loading the open end with a lumped capacitance. Physically this is accomplished somewhat as shown in Fig. 16-12. When the center conductor, which acts as the tuning plunger, is far from the

stepped-down cylindrical region, the structure can be considered a true quarter-wavelength cavity. When the center conductor protrudes into the small cylinder, the line is capacitively loaded, and the resonance that would normally occur at $3\lambda_0/4$ now occurs at a shorter wavelength or higher frequency, which gives a greater bandwidth without spurious response.

Often it is practical and economical to use higher modes of resonance than the basic one in wavemeters. This arrangement is commonly used in waveguide cavities, in frequency bands where nearby spurious resonances are not particularly inconvenient. It has been shown that resonance in a cylindrical waveguide cavity occurs when

$$f^2d^2 = A + Bn^2\left(\frac{d}{l}\right)^2 \tag{16-6-4}$$

where f = frequency, MHz
d = cylinder diameter, in.
l = cylinder length, in.
n = index of mode (number of half wavelengths along axis)
A = a constant depending upon mode 0.4781 for TE_{111} mode
B = a constant depending upon velocity of propagation, 0.34799×10^8 for air-filled cavity

The ability to adjust a cavity to resonate precisely with a given signal depends in part upon the Q of the cavity, which in turn depends upon the ratio of stored power to loss power in the cavity. In general, the higher the order of the oscillation mode in a cavity, the higher the Q is. For comparison, a coaxial cavity in x band (8.2 to 12.4 GHz) has a Q range from 1,500 to 4,000, while a TE_{111} mode resonator has a Q of about 10,000. The Q in some modes is as high as 100,000.

Coupling to Cavities. A little power from the signal source can be coupled either electrically or magnetically into the cavity of a frequency meter. In the first case, a short length of wire protrudes into the cavity at a point where the rf voltage on the wire couples well to the electric field. For magnetic coupling, a small loop carrying signal current is oriented to the resonant magnetic field of the cavity. The coupling is kept as low as possible to avoid degradation of the Q.

One still has the choice of making the wavemeter cavity either absorb part of the signal at resonance or transmit the signal only at resonance. The absorption (or reaction) arrangement is best considered a series resonant circuit connected across a transmission line. If the transmitted signal level as measured at signal frequency is varied in such a setup, a curve similar to that in Fig. 16-13 is obtained. In practice, if the frequency is constant, the frequency meter is tuned for maximum absorption. The kind of coupling to obtain an absorption curve is shown in Fig. 16-14. If a sweep generator is used as part of a test arrangement,

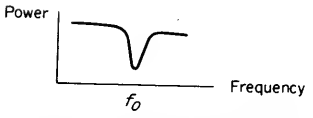


FIG 16-13 Tuning curve of a reaction or absorption frequency meter.

the absorption frequency meter can be left connected in the system as a variable frequency marker for an oscilloscope presentation.

For transmission coupling to a frequency meter, the arrangement of Fig. 16-15 is used. A matched input circuit feeds power into the resonant cavity, and at resonance, the cavity transmits power to the matched output circuit. Such a device can be considered a tunable bandpass filter with bandwidth dependent upon Q.

All types of frequency meters are customarily provided with a drive mechanism having negligible backlash and a calibrated frequency dial.

In summary, cavity frequency meters, depending upon the mode used in their resonant structure, can be classified by calibration accuracy and bandwidth freedom from spurious resonance. The following table shows a few types of cavity frequency meters and some of their typical specifications.

Type of cavity frequency meter	Bandwidth	Q	% Accuracy
TEM half wave	2:1	2,000–4,000	0.1
TEM quarter wave	3:1	2,000–4,000	0.1
TEM capacitively loaded quarter wave	4–10:1	500–4,000	0.1
TE_{111}	1.5:1	8,000–12,000	0.02–0.05
TE_{011}	1.1:1	50,000–100,000	0.005

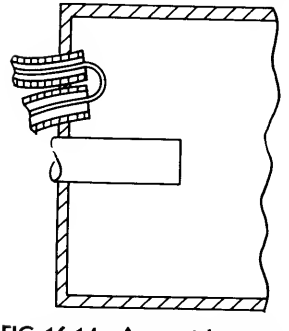


FIG 16-14 A coaxial arrangement for magnetic field coupling for reaction or absorption.

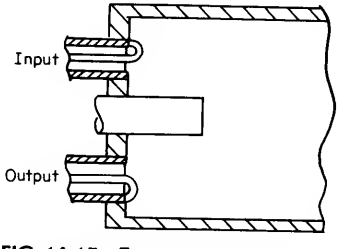


FIG 16-15 Transmission type of coupling in a coaxial structure.

Calibration accuracy can be impaired by mechanical wear, extreme temperature, and humidity. Better cavity types of frequency meter are designed with these factors in mind. Great efforts are made to select the correct materials to assure the least mechanical wear and at the same time achieve temperature compensation. Some frequency meters are also made of invar metal having practically zero thermal expansion. Humidity conditions can be met by good sealing or by correct surface preparations and drain-hole arrangements.

16-7 Spectrum Analysis

The most common way of observing a signal is to display it on an oscilloscope, with time as the x axis, as described in Chap. 11. This is a view of the signal in the *time* domain. It is also very useful to display signals in the *frequency* domain. This measurement method, often providing unique information unavailable, or practically unavailable, in the time-domain view, is called *spectrum analysis* and is the subject of this section. The instrument providing this frequency-domain view is the spectrum analyzer. On its CRT, the spectrum analyzer provides a

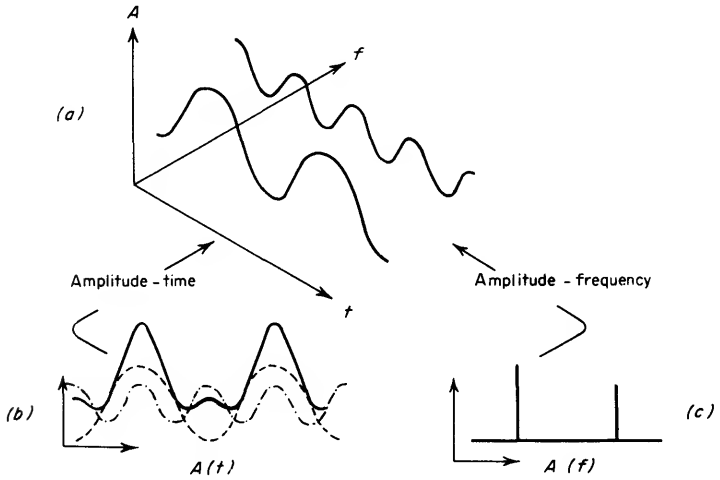


FIG 16-16 The time and frequency domains: (*a*) Three-dimensional coordinates showing time, frequency, and amplitude. The addition of a fundamental and its second harmonic is shown as an example. (*b*) View seen in the $A(t)$ plane. On an oscilloscope, only the composite $f_1 + 2f$ would be seen. (*c*) View seen in the $A(f)$ plane. Note how the components of the composite signal are clearly seen here.

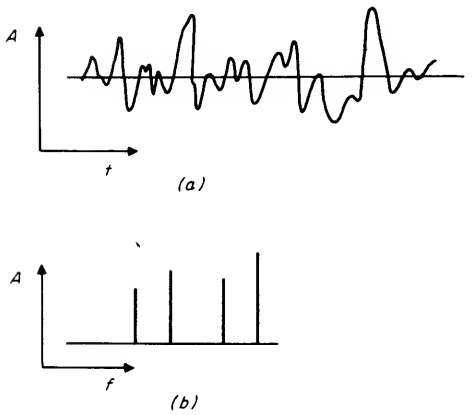


FIG 16-17 The time and frequency domains with incoherent signals.

calibrated graphical display, with frequency on the horizontal axis and voltage on the vertical axis. Displayed as vertical lines against these coordinates are the sinusoidal components of which the input signal is composed. From the height of each vertical line representing a component, its absolute amplitude is measured, and from its horizontal location, its frequency, just as the oscilloscope vertical and horizontal axes provide the amplitude and time dimensions of a signal.

Figure 16-16 shows how a simple signal $A(t)$ looks as viewed in the time domain on an oscilloscope (*b*) and in the frequency domain $A(f)$ on a spectrum analyzer (*c*).

Figure 16-16 shows the analysis of a complex but repetitive time-domain signal, a signal that was generated, so to speak, in the time domain. In this case, the frequency-domain components are coherent, that is, are exact frequency multiples and bear a definite phase relationship to each other. In another important type encountered in spectrum analysis several sinusoidal components of differing frequency that are unrelated to each other are present. These components could be signals from unrelated oscillators, which could be carrier channels of a telephone of a telemetry-frequency multiplex system or perhaps several radio stations that are simultaneously present. These signals are frequency-domain signals; they can only be distinguished from each other in the frequency domain. Figure 16-17 illustrates this; the amplitude and frequency of each component in the frequency-domain view are clearly visible. However, no coherent picture results from the time-domain view.

Time-domain analysis and spectrum analysis of signals complement

each other. Each has its own area of almost exclusive application, and there is a gray area where both types of measurement yield important information. The digital computer operates almost exclusively in the time domain. Also in the time domain is the time-multiplexed pulse-code-modulation telephone communication system, where each telephone channel is sampled sequentially and each sample level digitized into a digital word. The transmitted signal is then a series of digital words each representing the level of a particular channel sampled at a particular time. Signals from oscillators, mixers, and modulators are primarily frequency-domain signals.

In the rf and microwave-frequency range, many relatively narrow band signals spread over a wide frequency range are often encountered. An example is the transmitted signal in a frequency-multiplexed telephone communications system. In one instance each of 3,600 telephone channels is put on its own carrier, which is separated from its neighbors by 4 kHz in frequency. The frequency spread is from 100 kHz to 15 MHz. Critical measurements here are the frequency and amplitude of the individual channel components, which the spectrum analyzer can give. The amplitude axis on the spectrum analyzer can be made logarithmic, which enables components that differ very greatly in amplitude to be displayed simultaneously. The oscilloscope cannot distinguish this information on individual channel components.

Although capable of covering a wide frequency range, such as from 0 to 1,000 MHz, the spectrum-analyzer detection is narrow band, making it a very sensitive detector of narrow-band signals easily capable of measuring a sine wave at the microvolt level. The oscilloscope is a broadband instrument and so is sensitive in detecting broadband signals such as narrow pulses but not narrow-band signals such as a low-level sine wave. The spectrum analyzer is a considerably more sensitive detector than other broadband detectors with wide frequency range such as rf voltmeters, power meters, and crystal detectors.

Figure 16-18*b* shows how the spectrum analyzer can measure extremely small amounts of distortion, far beyond the capability of an oscilloscope, by measuring very small harmonic components, generated from a sinusoidal test signal by the device under test. However, Fig. 16-18*a* shows how at higher levels the oscilloscope can show where in the test-signal cycle the distortion is occurring, which can lead to uncovering its cause. Spectrum-analyzer applications include the measurement of signal level and the measurement of frequency and its response, harmonic and intermodulation distortion, frequency stability, spectral purity, modulation index, and attenuation.

Scanning and Real-time Spectrum Analyzers. There are two types of spectrum analyzers: scanning types, which scan in frequency, and nonscanning

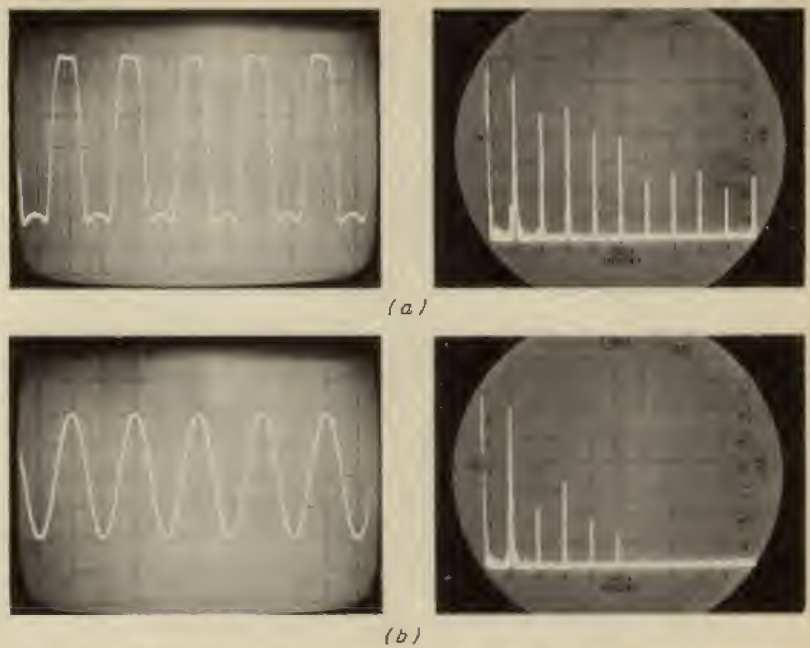


FIG 16-18 Overdriving an amplifier results in a severely distorted waveform easily observed with the oscilloscope; however, quantitative measurements of distortion levels are difficult to obtain. The scope calibration is 0.2 V per division vertically, 0.05 μsec per division horizontally. When the input power level is reduced (*b*), the distortion is hardly observable. The spectrum analyzer easily gives quantitative information about the distortion of the two signals. The frequency scale is 10 MHz per division centered at 50 MHz, and the reference level is +20 dBm. (The response at the far left is the zero frequency indicator.) The 10-MHz signal input is at −30 dBm in *A* and at −40 dBm in *B*. Since the normal amplifier gain is 40 dB, gain compression is about 7 dB in *A*. Second harmonic distortion is reduced from 14 dB down to 38 dB down by the 10-dB reduction in signal input. The effect of input signal level on the other harmonics is also easily discerned.

types, also called *real-time spectrum analyzers*. The scanning types are essentially swept receivers, both superheterodyne and tuned rf (trf), whose tuning is electrically swept over the frequency range of observation by a scanning signal that also controls the horizontal position of the spot on the CRT ray tube. There are two kinds of real-time, nonscanning spectrum analyzers: the multichannel spectrum analyzer, and the computational spectrum analyzer or Fourier analyzer. These real-time analyzers "look" over all parts of their frequency display range simul-

taneously and so present the spectrum of an electrical event as soon as it happens. The scanning spectrum analyzer can only "look" at a single frequency at a given instant. At the present time, the most important type of spectrum analyzer at rf and microwave frequencies is the swept superheterodyne. This type and its applications will be the main subject of this chapter, although the other kinds will be described and some of the strengths and weaknesses of the various kinds discussed.

16-8 The Swept Superheterodyne Spectrum Analyzer

Figure 16-19 shows the controls and display of a typical rf or microwave swept superheterodyne spectrum analyzer. The horizontal axis, frequency, is linear, with the center frequency given by the tuning dial on the instrument which is readable to within about 1 percent. The accuracy of the frequency scan width of the display is about 5 percent. The use of accessories, which can greatly extend the ability of the spectrum analyzer to measure frequency accurately, will be discussed later. In a microwave instrument, the horizontal axis can display as wide a range as 2 to 3 GHz, for a broad survey, to as narrow as 30 kHz, for a highly magnified view of any small portion of the spectrum. Signals, at microwave frequencies and separated by only a few kHz, can be seen individually. The frequency range covered by the instrument is from 1 MHz to 40 GHz. In an rf instrument covering from 1 kHz to 100 MHz, the spectrum window ranges from 100 MHz down to 2 kHz and sometimes less, with signals that are 100 and even 10 Hz apart capable of separate identification. The vertical axis of the display is logarithmic, 10 dB per division, and capable of displaying a 70-dB range. Linear display can also be used. The amplitude axis is calibrated, which enables the measurement of absolute signal level to within about ½ dB at rf, 1 dB at low microwave frequencies (2 GHz), and 2 to 3 dB at x band (8 to 12.4 GHz). Input impedance is generally 50 Ω. Sensitivity is such that signals of -110 dBm (≈ 1 μV) can be measured at rf, -90 dBm (≈ 10 μV) at 1 GHz, and -75 dBm (≈ 50 μV) at 12.4 GHz. Input attenuators permit input levels to $+20$ dBm (2 V).

Theory of Operation. Figure 16-20 is the basic block diagram of a spectrum analyzer covering the range of 500 kHz to 1 GHz, which is representative of the superheterodyne type. The input signal is fed into a diode mixer, which is driven to saturation by a strong signal from the local oscillator. This oscillator is linearly tunable electrically over the range of 2 to 3 GHz. The input mixer multiplies (heterodynes) the input signal and local oscillator signal together and so provides two signals at its output that are proportional in amplitude to the input signal but of

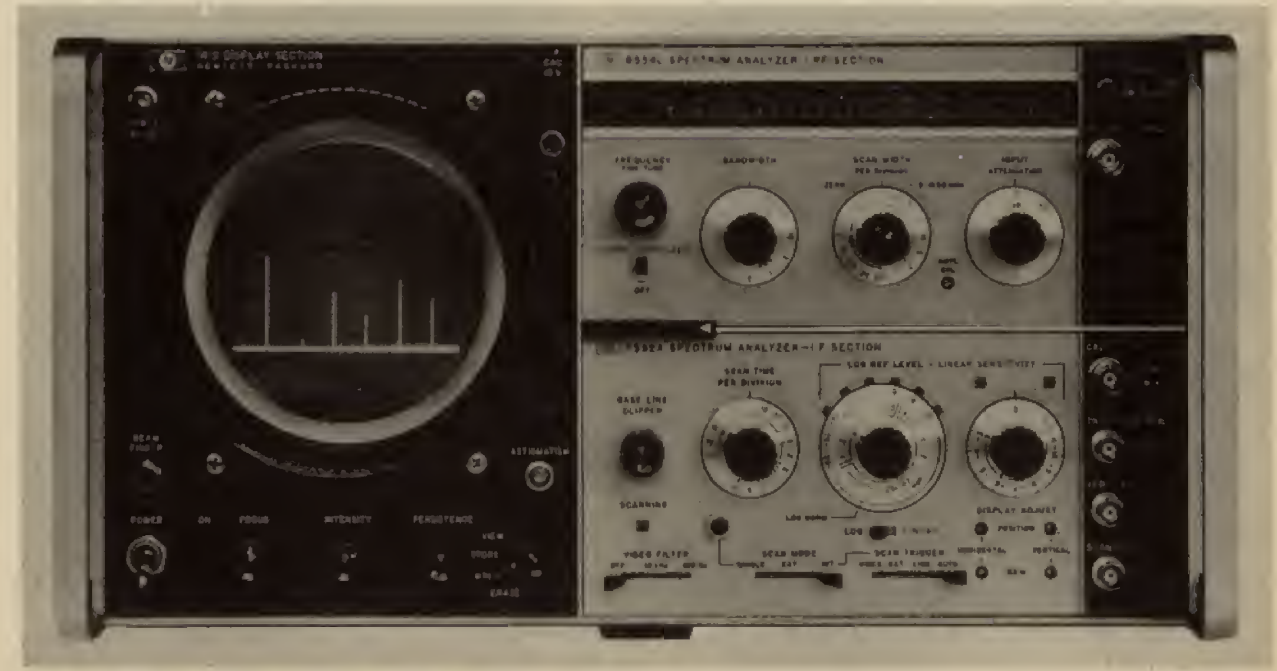


FIG 16-19 Swept superheterodyne spectrum analyzer.

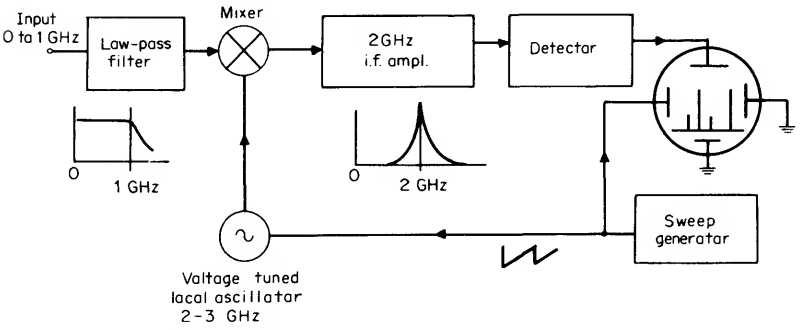


FIG 16-20 Swept superheterodyne spectrum analyzer covering 50 kHz to 1 GHz.

frequencies that are the sum and difference between the frequencies of the input signal and the local oscillator signal.

The intermediate-frequency amplifier is tuned to a narrow band around 2 GHz. As the local oscillator is tuned over the range from 2 to 3 GHz, only input signals that are separated from the local-oscillator frequency

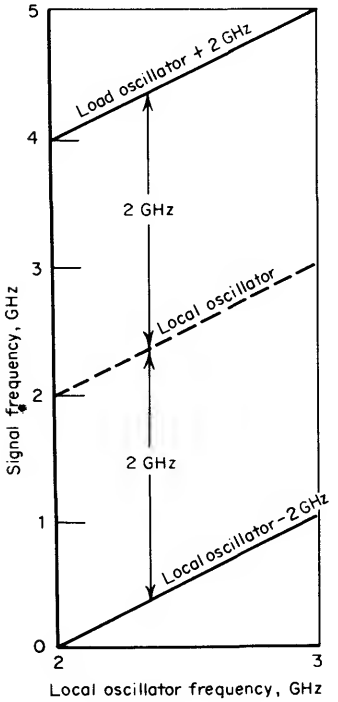


FIG 16-21 Superheterodyne tuning chart.

by 2 GHz will be converted to the intermediate-frequency band, pass through the intermediate-frequency amplifier, be rectified in the detector, and produce a vertical deflection on the CRT. Figure 16-21 plots the input signal frequencies to which the spectrum analyzer will respond as a function of local-oscillator frequency. From this it is seen that as the sawtooth signal sweeps the local oscillator linearly from 2 to 3 GHz, the tuning of the spectrum analyzer as a receiver is swept linearly from 0 to 1 GHz. The sawtooth scanning signal is also applied horizontally to the CRT to form the frequency axis of the display, the left edge representing zero frequency and the right edge 1 GHz. Any horizontal driving waveform could be used for the scanning function, but a sawtooth provides a linear time axis useful in certain applications. Notice in Fig. 16-21 that the spectrum analyzer will also be sensitive to signals from 4 to 5 GHz, referred to as the image frequency of the superheterodyne. A low-pass filter with a cutoff a little above 1 GHz at the input suppresses these spurious signals.

Figure 16-22 is a more detailed block diagram of the spectrum analyzer just described. The frequency of the first local oscillator is controlled with a selectably attenuated signal from the scan generator combined with an adjustable bias level from the center frequency control in the voltage control block. The attenuator controls the frequency axis calibration of the display by controlling the frequency range over which the first local oscillator is scanned. The adjustable bias level sets the frequency about which the local oscillator is scanned and thus the center frequency of the display. This arrangement gives a magnified display, expandable over a wide range, of any portion of the spectrum in the frequency range of the instrument.

Several conversions are used in the intermediate-frequency amplifier chain, which ultimately gets down to a 3-MHz intermediate-frequency amplifier having a bandwidth of a few hundred hertz to provide this high selectivity over the whole range of the instrument. The high first intermediate frequency is necessary for wide "image" separation, the low last intermediate frequency is necessary for narrow-band filtering unobtainable at the first intermediate frequency.

Frequency Resolution and Bandwidth. Frequency resolution is the ability of the spectrum analyzer to separate signals closely spaced in frequency. Two factors determine resolution, the bandwidth or selectivity of the intermediate-frequency amplifier and the frequency stability of the spectrum analyzer, as determined by the drift, residual FM, and phase noise of the local oscillators. The scanning action in the spectrum analyzer slides the input spectrum past the final intermediate-frequency amplifier filter. This can also be viewed as sliding the intermediate-frequency filter characteristic past the input spectrum as the analyzer scans. Because of this, the magnified display of a single-frequency continuous-wave signal

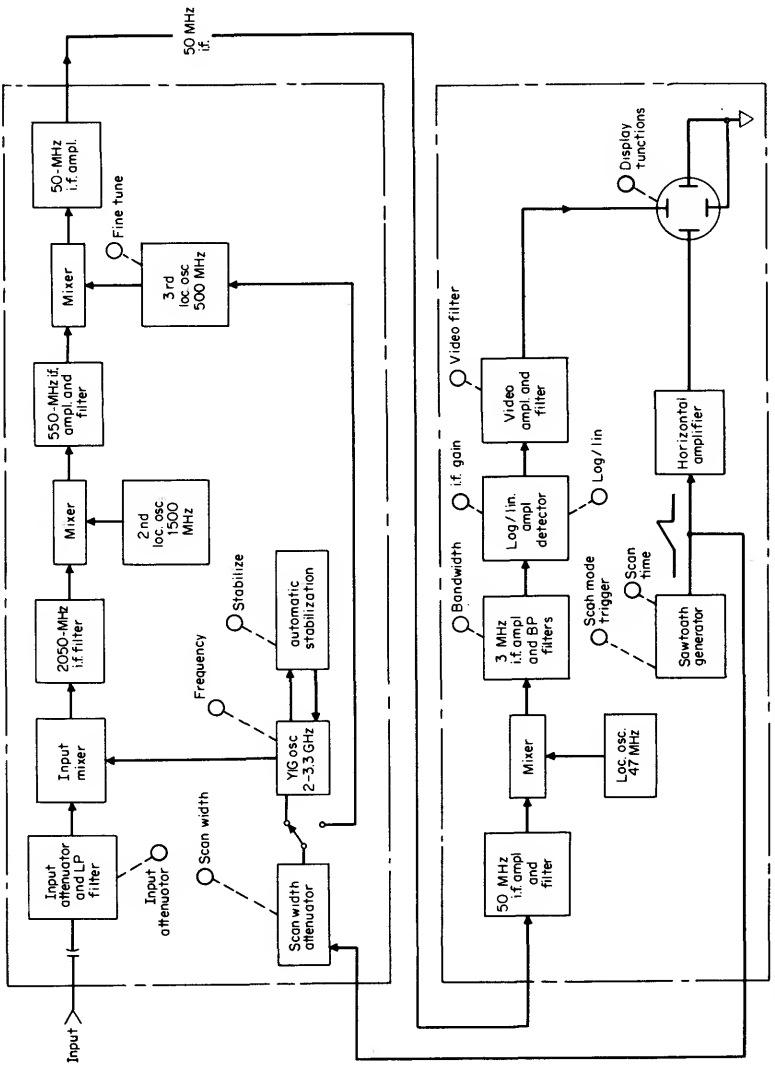


FIG 16-22 More details of swept superheterodyne spectrum analyzer.

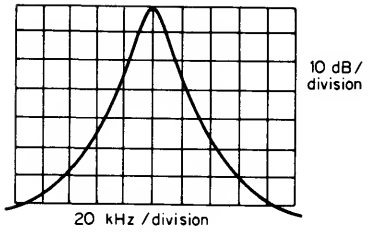


FIG 16-23 Selectivity characteristic for a 3-db bandwidth of 10 kHz.

is a plot of the frequency selectivity characteristic of the intermediate-frequency amplifier filter. Figure 16-23 shows this.

The intermediate-frequency filter characteristic is the window through which the instrument "looks" at separate signals. To resolve two signals and measure the amplitude and frequency of each, each must appear in the window separately. Two continuous-wave signals with a separation of less than the intermediate-frequency bandwidth would both be in the passband at the same time and could not be distinguished. Multiple synchronously tuned intermediate-frequency filters are used. These approach a gaussian response, which will not ring and produce outputs that could be misinterpreted when a signal is swept rapidly through them or in an impulse-input situation.

Figure 16-23 shows the selectivity characteristic of a typical filter. The ability to resolve signals differing widely in amplitude is determined by the skirt characteristic, since the skirt resulting from the large signal can mask the smaller signal. Figure 16-24 shows that the filter shown in Fig. 16-23 can resolve two equal-amplitude signals 20 kHz apart, but signals differing by 60 dB would have to be more than 70 kHz apart to be separated. A wide selection of bandwidths is provided in the spectrum analyzer, to avoid sweep desensitization and provide greater sensitivity with impulse inputs, as will be described later.

The other factor determining resolution is the frequency stability of the spectrum-analyzer local oscillators. These oscillators must have

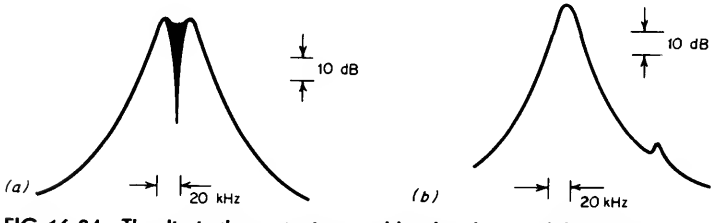


FIG 16-24 The dip in the center is caused by the phasing of the two signals to cancel each other. (*a*) Two equal-amplitude signals, 20 kHz apart; (*b*) Two signals of 60-dB difference in amplitude and 70 kHz apart.

greater absolute stability than any signals that are to be analyzed and resolved. Residual FM in the spectrum analyzer will smear the display, and phase noise will add noise skirts to the filter skirts, both reducing resolution as shown in Fig. 16-25.

Sweep Desensitization. Sweep desensitization is an effect, caused by scanning a spectrum analyzer too fast, which results in loss of amplitude calibration, sensitivity, and resolution. It is easily detected and corrected if understood. During scan the signal must remain in the bandpass of the intermediate-frequency filter long enough to allow the amplitude of the signal in the filter to build up to the proper value. A simple rule of thumb for avoiding sweep desensitization is that the scan velocity in hertz per second must not exceed the square of the 3-dB bandwidth of the intermediate-frequency filter in hertz. The *frequency* resolution of the spectrum analyzer is also reduced by this effect because the displayed time-domain transient response of the filter masks its frequency-selectivity characteristic. Figure 16-26 shows the extent of these effects accurately calculated for a gaussian filter.

Three controls of the spectrum analyzer are involved with this effect: the *bandwidth* (resolution), the *scan width* (controlling the frequency span of the display), and the *scan time.*

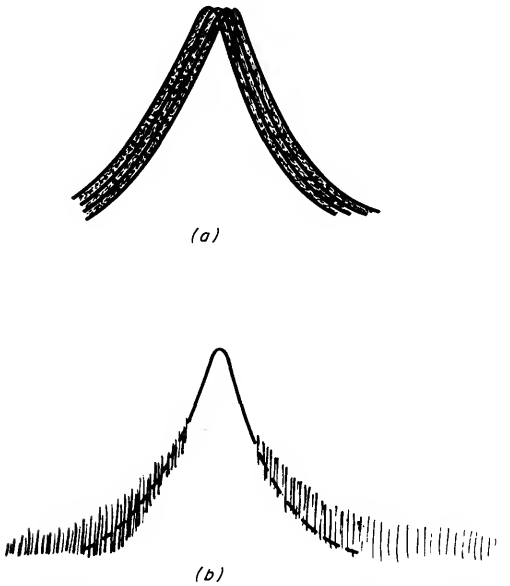


FIG 16-25 Effects on resolution in a spectrum analyzer of (*a*) residual FM and (*b*) noise sidebands.

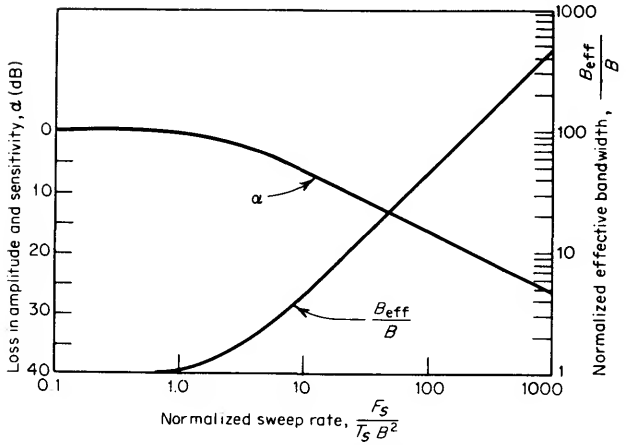


FIG 16-26 Sensitivity loss and normalized effective bandwidth versus normalized sweep rate; F_s **= sweep width,** T_s **= sweep time,** B **= 3-dB intermediate-frequency bandwidth, and** B_{eff} **= effective bandwidth.**

Sensitivity The ability of the swept superheterodyne spectrum analyzer to measure small signals is determined by its own internally generated noise. Typical noise figures vary from 25 dB at low frequency to 40 dB at 12 GHz. The internally generated noise referred to the spectrum-analyzer input exceeds basic thermal noise by these noise figures. At room temperature, the thermal-noise power spectral density 4×10^{-9} μW/MHz, or −114 dBm in a 1-MHz bandwidth. The available thermal noise power, in watts,

$$P_H = kTB \tag{16-8-1}$$

where B = bandwidth of system
k = Boltzmann's constant, 1.38×10^{-23}W-sec/°K
T = absolute temperature, °K

The noise on the spectrum-analyzer display is that contained only within the passband of its intermediate-frequency filter. Although the spectrum analyzer covers a wide frequency range (by sweeping) it is a narrow-band instrument and therefore very sensitive to continuous-wave signals. Noise power is proportional to bandwidth, and so the highest sensitivity to continuous-wave signals is obtained by using the narrowest bandwidths. Figure 16-27 shows the levels of thermal noise for various bandwidths at room temperature.

From the noise figure and the thermal noise level in various bandwidths, the ability of the spectrum analyzer to measure small signals

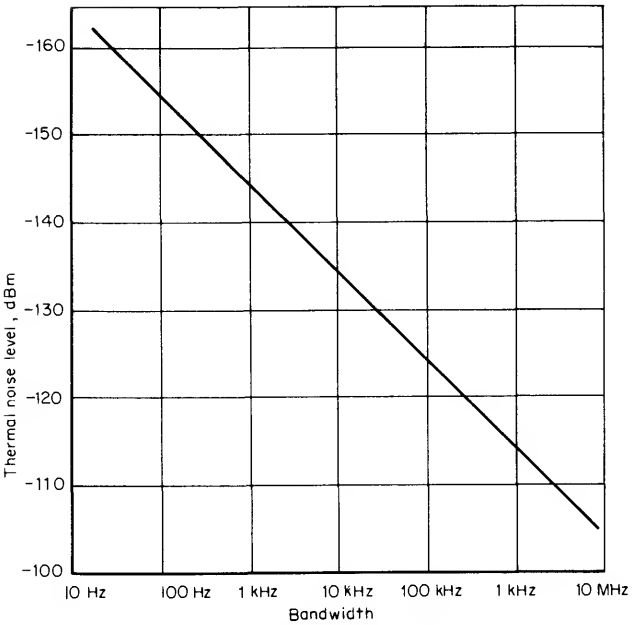


FIG 16-27 Thermal noise level versus bandwidth.

can be calculated. For instance, with a 100-kHz bandwidth and a 25-dB noise figure, the analyzer noise level is -124 dBm $+$ 25 dB $=$ -99 dBm. To be easily seen above the noise "grass" level on the display, a continuous-wave signal should be 10 dB above the spectrum analyzer's own noise level, or -99 dB $+$ 10 dB $=$ -89 dBm, or 8 mV across 50. Figure 16-28 shows the appearance of a signal that is 10 dB above the noise power level of a spectrum analyzer with 100-kHz bandwidth.

Broadband preamplifiers with low noise figures, lower than that of the spectrum analyzer, can be used to improve sensitivity. The net noise figure of cascaded blocks is

$$N = N_A + \frac{N_B}{G_A}$$

where N = system numeric noise figure
N_A = numeric noise figure, first block
N_B = numeric noise figure, second block
G_A = numeric power gain, first block (10 dB = power gain of 10); numbers are numerics, not decibels, that is, 100, not 20 dB

FIG 16-28 Appearance of a signal that is 10 dB above the noise power level of the spectrum analyzer.

Transistor, tunnel-diode, and traveling-wave-tube amplifiers having 5- to 10-dB noise figures, with gains of 20 to 30 dB and continuous coverage over wide frequency ranges are available. If the gain of the amplifier is greater than the noise figure of the analyzer, the system noise figure is that of the amplifier. Otherwise, providing the amplifier noise figure is much less than that of the analyzer, the system noise figure will be the analyzer noise figure less the gain of the amplifier, all in decibels.

Dynamic Range. Spectrum analyzers usually provide a selection of *linear*, proportional to voltage, *squared*, proportional to power, or *logarithmic* modes of amplitude display. The logarithmic mode, with its ability to display signals that differ greatly in amplitude, is the most frequently used. In the spectrum analyzer described earlier, the logarithmic vertical response is accomplished by a successively limiting final intermediate-frequency amplifier with a logarithmic response over a 75-dB range. Figure 16-18 shows the 10-dB per division logarithmic presentation of the spectrum analyzer and the simultaneous display of 300 mV and 300 μV components.

The dynamic range of the spectrum analyzer expresses its ability to display the true spectra of large and small signals simultaneously. With signal levels within the dynamic range of the instrument, spurious signals, which result from the distortion of the large signals by the analyzer itself and could either mask the small signals or erroneously appear as small signals, will not appear on the display. The dynamic range that is free of spurious signals can be defined as the ratio of the signal level to the noise level at signal levels where spurious distortion products just

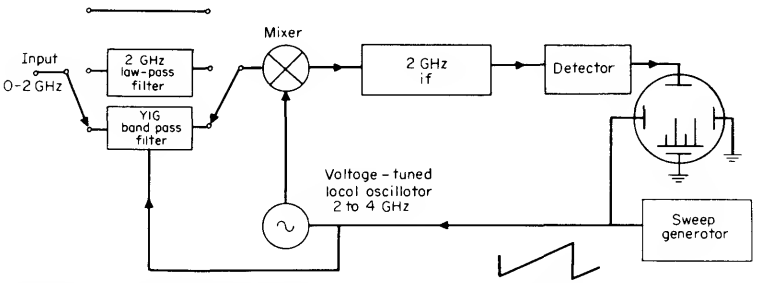


FIG 16-29 Harmonic mixing in swept superheterodyne spectrum analyzer from 0 to 12 GHz.

begin to become visible above the noise level of the display. Present spectrum analyzers have 60- to 70-dB dynamic ranges, free of spurious signals, for continuous-wave signals except at very high frequency. Input attenuation can adapt the analyzer's dynamic range to higher signal levels.

Harmonic Mixing. Harmonic mixing is used to extend the frequency range of the swept superheterodyne spectrum analyzer. In the spectrum-analyzer example described earlier, the input signal was mixed with the *fundamental* frequency of the first local-oscillator signal. Figure 16-29 is the block diagram of a spectrum analyzer with the same configuration as the previous one but with a first local-oscillator range of 2 to 4 GHz. The instrument has fundamental and harmonic mixing to cover from 10 MHz to above 12 GHz. Figure 16-30 is a plot of the transfer function $A(t)$ of the input mixer, which shows the diode gate being switched

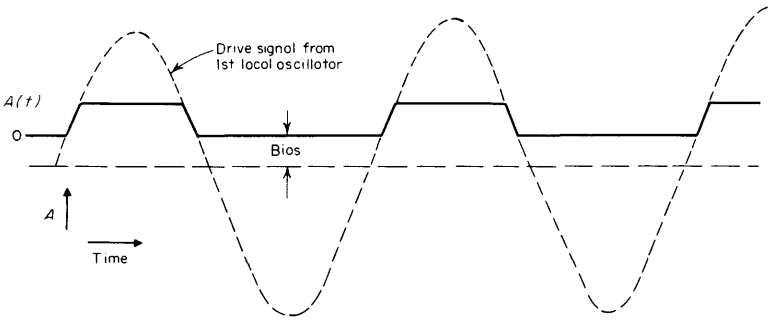


FIG 16-30 Transfer function of input mixer of a spectrum analyzer, $A(t) = A(t) = A_0 + A_1 \sin(\omega_0 t) + A_2 \sin(2\omega_0 t + \phi_2) + A_3 \sin(3\omega_0 t + \phi_3) + \cdots$

off and on by the first local-oscillator drive signal. This transfer function of time is shown expressed as a Fourier series function of frequency. Bias is applied to the drive waveform to unbalance the duty cycle so that even as well as odd harmonics are present. An input signal is multiplied by this transfer function, which produces sum and difference frequency output signals with each term of the Fourier series. In other words, the input signal can be heterodyned with the fundamental, second harmonic, third harmonic, etc., of the first local-oscillator frequency. Any output signals produced at 2 GHz will pass through the system and be displayed as already described. So the analyzer will respond to signals which differ not only from the fundamental by 2 GHz, but also by 2 GHz from the second harmonic, third harmonic, etc. Figure 16-31 plots these input signal responses versus the first local-oscillator fundamental frequency. Conversion loss increases with the order of the mixing mode N, and the conversion gain decreases as approximately $1/N$.

Harmonic mixers respond to several input frequencies simultaneously, but preselection filters can be used to eliminate the confusion. Bandpass broadband filters can help with higher-order harmonic mixing modes, but the most effective solution is a tracking narrow-band filter which can be adjusted to track a desired harmonic mixing mode as the spectrum analyzer is tuned and scanned. Such a filter is the YIG filter, consisting of one or more coupled yittrium-iron-garnet resonators whose resonant frequency is proportional to the strength of the field from an electromagnet in which they are placed. This filter can be biased and electrically swept to allow only the signal matching a desired mixing mode to enter the input mixer. This is, of course, like the tuned front end on an ordinary broadcast receiver.

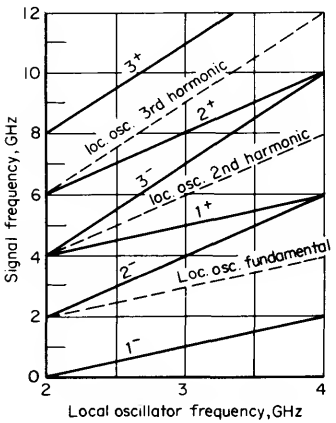


FIG 16-31 Wide-range conversion scheme with harmonics of the local oscillator.

16-9 Wave Analyzers

A wave analyzer is a tuned voltmeter with a very narrow passband. It is a narrow-band superheterodyne receiver like the spectrum analyzers previously described, but with a meter readout instead of a CRT, and it usually must be manually tuned through the frequency range being examined. The wave-analyzer tuning is manually centered on the frequency of an input signal, and then the signal amplitude is read out on the meter. Wave analyzers are used in the low rf ranges, below 50 MHz and down through audio, and they provide very high resolution of frequency. To make tuning less critical, wave analyzers have intermediate-frequency filters with flat tops and very steep skirts. Steep skirts can be tolerated, for transient response is not important because there is no swept display. On the contrary, the spectrum analyzer does not need a flat-topped selectivity characteristic, for the peak of the response will always be swept through the signal. Figure 16-32 shows the shapes of the two different kinds of selectivity characteristics.

A wave analyzer does not provide the broad and instant view of a spectrum range that a spectrum analyzer does and this can be a disadvantage. Very often something in one part of the spectrum is affecting something else in another, and a view of all parts is valuable. However, a scanning spectrum analyzer with comparable resolution with that of the wave analyzer would have to sweep so slowly that the complete display would be slow to build up and the "instant view" would not be available anyway. Tuning and adjusting the display are difficult under these circumstances. Some wave analyzers can be slowly swept and their outputs recorded on an *xy* recorder to provide a broad spectral view.

Wave analyzers are available that can measure signal frequency very accurately. Electronic counters are built in to measure local-oscillator frequencies and from them to compute and present the center frequency of the wave-analyzer passband in a 5- and even 7-place digital readout.

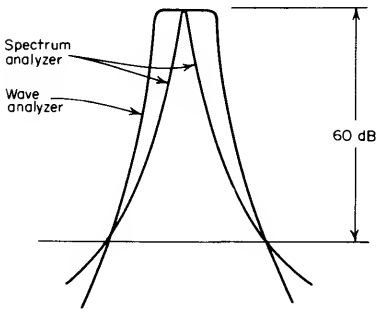


FIG 16-32 Selectivity characteristics of the spectrum analyzer and wave analyzer compared.

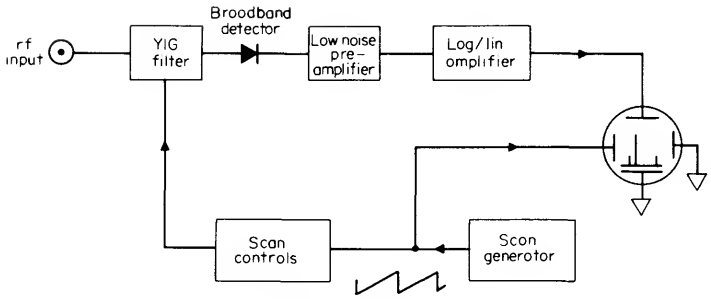


FIG 16-33 A yig-tuned trf spectrum analyzer.

The advantages over the direct use of a counter are that the frequency of very small signals can be measured and a signal can be selected out of many even larger signals that might be present. Wave analyzers usually have automatic frequency control, in which the tuning automatically locks to a signal and so makes it possible to measure the amplitude of signals that are drifting in frequency by amounts that would carry them outside the widest passband available.

16-10 The TRF Spectrum Analyzer

The heart of the trf spectrum analyzer is a narrow-band input filter that is sweepable over the frequency range of the instrument. Figure 16-33 is a block diagram of the trf analyzer. Analysis is accomplished by scanning the input filter across the spectrum in proportion to the CRT horizontal deflection, the detector output producing the CRT vertical deflection. This type of analyzer has been made practical by the electrically tunable yig filter, already described as a preselector for the harmonic-mixing superheterodyne spectrum analyzer. Consequently, this type of analyzer is used presently only in the microwave frequency range, 1 to 18 GHz, of the yig filter.

The trf analyzer has three advantages: simplicity, the ability to display very wide spectra such as a single display of the whole 1- to 18-GHz frequency range, and freedom from the multiple and spurious responses that can occur with the superheterodyne analyzer. Its disadvantages are very low resolution of frequency and poor sensitivity compared with those of the superheterodyne. The selectivity is that of the microwave yig filter, so it has 20-MHz bandwidth compared with as low as 1 kHz for a superheterodyne in the same frequency range. Because of lack of intermediate-frequency amplification, sensitivity is about -45 dBm compared with -100 dBm for the superheterodyne. Present application is limited to relatively large signals spaced widely in frequency.

16-11 Multifilter Real-time Spectrum Analyzer

A real-time spectrum analyzer presents the effect of changes in all input frequencies on its spectrum display as soon as they occur. The final limitation to the speed of presentation is the degree of frequency resolution with which the effect of the change is to be seen on the display. That is, this limitation is the inability to see the spectral effect of a change with a given frequency resolution in a time shorter than the rise time corresponding to that given bandwidth. The 10 to 90 percent of rise time of a bandpass device is approximately 1/(3-dB bandwidth), and so the minimum time to see a change with 10 Hz, 3-dB resolution is 0.1 sec. To resolve changes in two components which have the same 10-Hz frequency separation but differ more greatly in amplitude takes longer. Real-time analysis has important merits for making measurements where the frequency resolution is high, such as at audio frequencies and below. The very slow scan rate necessary to avoid sweep desensitization and deresolution in a swept superheterodyne spectrum analyzer with narrow bandwidths is a severe handicap under these circumstances. For instance, a display 1-kHz wide with a resolution of 10 Hz requires a display rate of 10 sec per display. The time lag between making an adjustment to a circuit (or even to the analyzer) and the display presentation makes

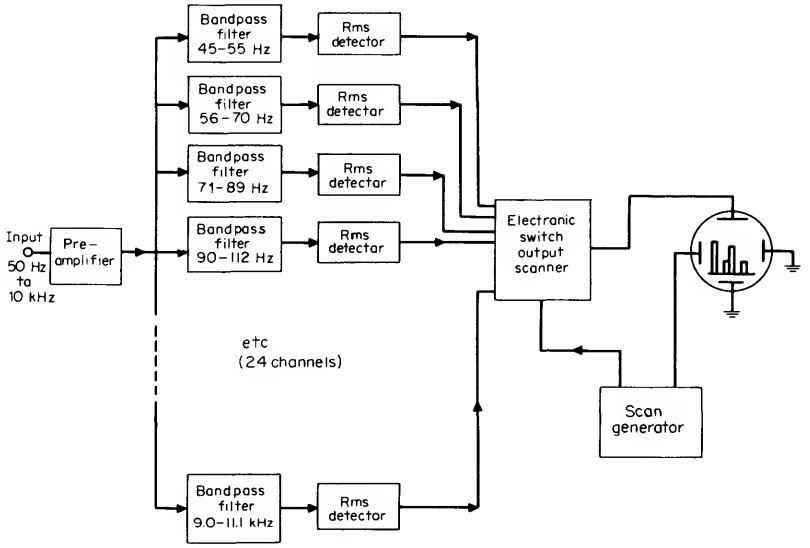


FIG 16-34 Multifilter real-time spectrum analyzer.

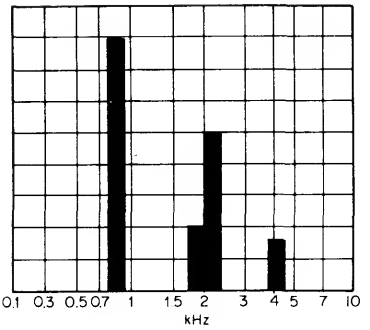


FIG 16-35 Multifilter real-time spectrum-analyzer display.

these operations awkward. A very important benefit of real-time analyzer techniques is to increase the display rate up to at least 3 Hz, which is about as fast as a human operator needs for the adjustment observation sequence. The rate may be up to 20 Hz, where display flicker will not be noticeable with normal phosphors.

If the swept superheterodyne spectrum analyzer is viewed as stepping across its display in increments equal to its bandwidth, it must dwell in each slot for at least the rise time corresponding to its bandwidth to avoid sweep desensitization and deresolution as previously described. In effect, the bandwidth slots in the spectral display are examined serially. The multichannel spectrum analyzer gets its speed by examining these slots in parallel. Figure 16-34 is the block diagram of an example of this type of analyzer that covers from 50 Hz to 10 kHz. Figure 16-35 shows the display of a signal on this type of analyzer.

The disadvantages to this type of spectrum analyzer are complexity, poor resolution, inflexibility of the frequency axis of the display, and inability to adjust resolution or view magnified portions of the display in detail. Real-time spectrum analyzers are not at present used at rf or microwave frequencies where high frequency resolution is not required.

16-12 *The Tracking Generator Counter*

If a signal generator can be made to accurately and automatically track the frequency to which a wave analyzer or spectrum analyzer is tuned, the resulting instrument system is valuable in measuring the frequency response of a network. As the analyzer is scanned over the frequency range of interest, the generator automatically produces the correct frequency at any time. Now, if a digital frequency counter is added and connected to the generator terminals, a very accurate frequency scale is established. Figure 16-36 shows how the tracking signal is generated.

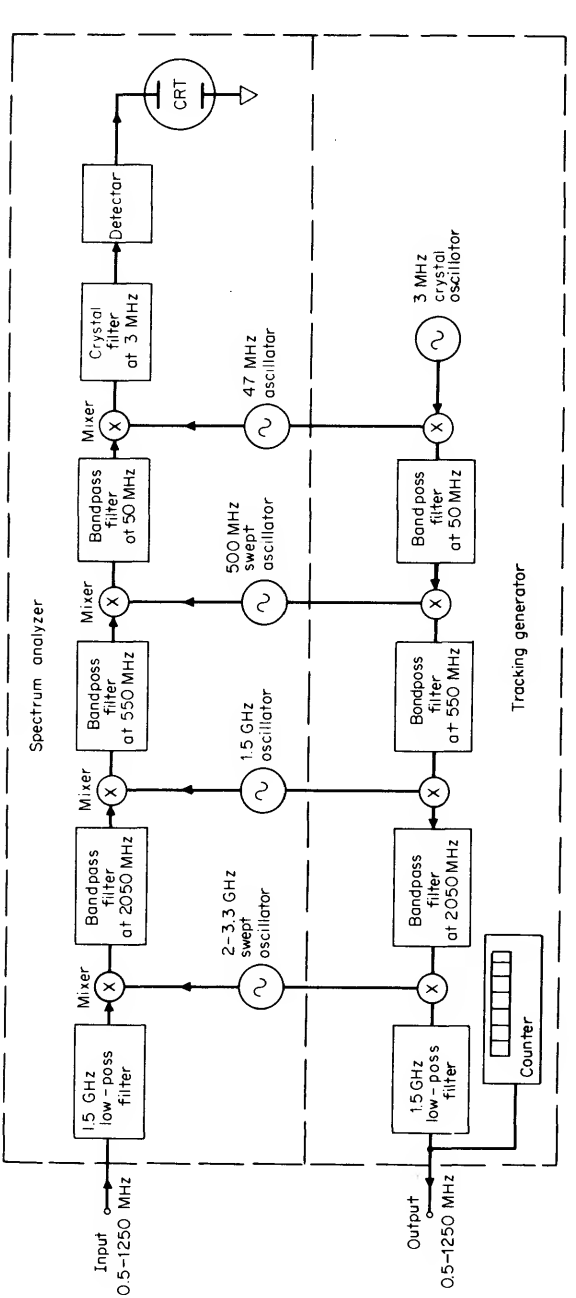


FIG 16-36 Tracking generator counter.

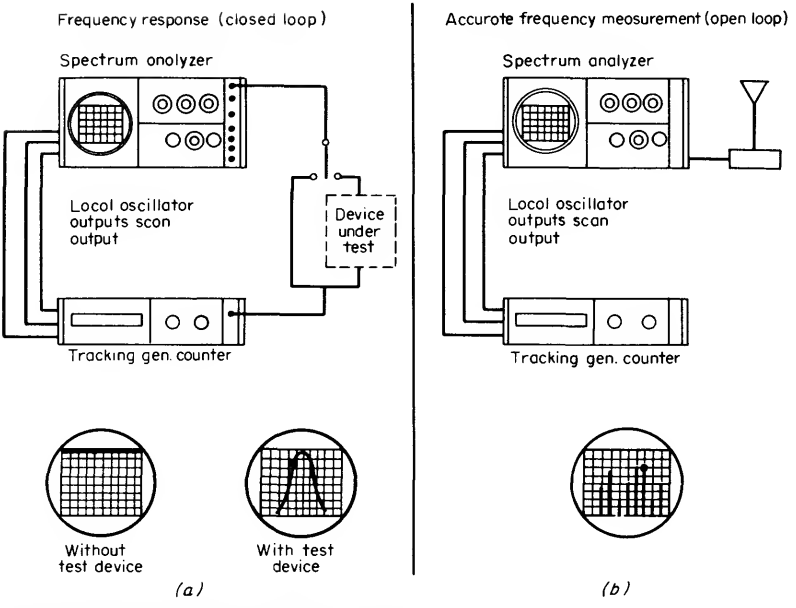


FIG 16-37 Tracking generator counter applications. Intense marker indicates where counter gives the frequency.

A crystal-controlled oscillator signal at the same frequency as the spectrum-analyzer final crystal filter is mixed with local oscillators as shown, which ultimately produces a signal that will track the spectrum-analyzer passband even on its narrowest bandwidth.

Figure 16-37 shows the frequency-response measurement setup. The tracking generator signal is connected to the input of the test network and the spectrum analyzer connected to the output. The amplitude frequency response of the network is plotted on the spectrum-analyzer CRT. The frequency response can be determined over very wide range in amplitude, over 100 dB. This is due to the high sensitivity of the analyzer and the advantage that this tracking, tuned system is not sensitive to harmonics of the test signal that may be present or generated.

16-13 Techniques and Applications for Analyzers

From the discussions above, many of the ways of using spectrum analyzers are obvious. To measure the amplitudes of the frequency components of a signal, after the center frequency, scan width, and other

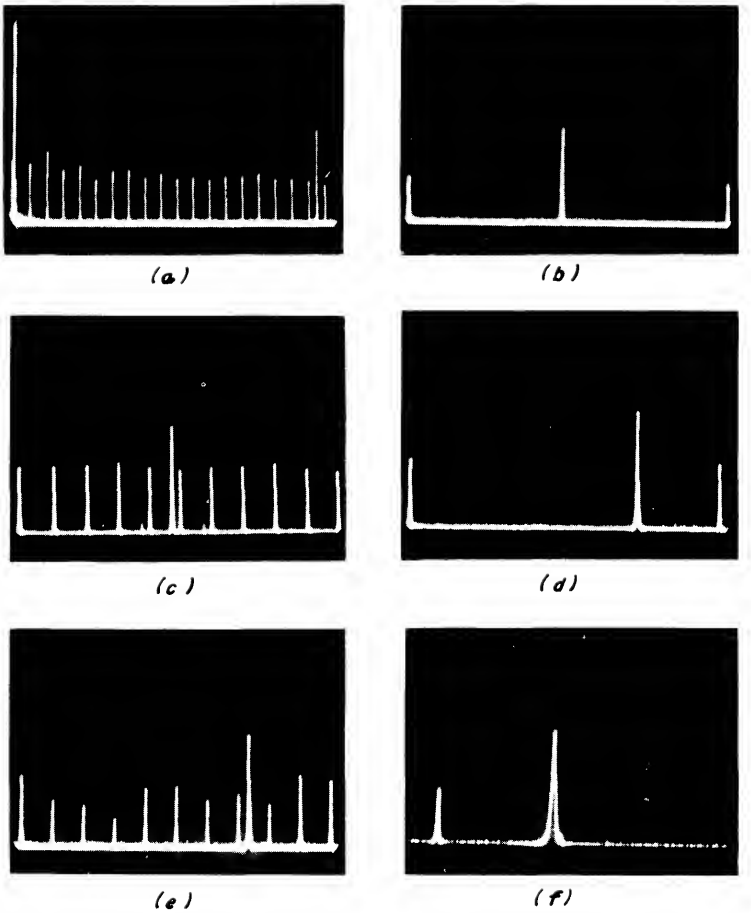


FIG 16-38 Accurate determination of a signal frequency on a spectrum analyzer.

panel adjustments have been made to produce a satisfactory display, one simply observes their heights on the CRT screen. Calibration can usually be chosen in dBm (0 dBm = 1 mW) or volts. Frequencies can be observed from the horizontal positions of the spectral lines, or for greater accuracy, signals with accurately known frequencies can be superimposed upon the measured signals. Digital displays are available.

The photographs in Fig. 16-38 show how the frequency comb generator can be used to improve the accuracy of frequency determination

with a broadband spectrum analyzer. Figure 16-38*a* shows the analyzer display of a signal combined with the 100-MHz comb. The large spike at the left is caused by local-oscillator feedthrough in the spectrum analyzer and provides a convenient zero-frequency reference. Counting the comb frequency components from the left shows that the signal lies between 1,800 and 1,900 MHz. Now the analyzer is tuned to place the 1,800-MHz marker at 0 cm and the analyzer-spectrum width is set to 10 MHz/cm (Fig. 16-38*b*). Switching to the 10-MHz comb and again counting harmonics shows that the signal is between 1,840 and 1,850 MHz (Fig. 16-38*c*). The spectrum width is next switched to 1 MHz/cm (Fig. 16-38*d*) and the 1-MHz components are added to the 10-MHz comb, as in Fig. 16-38*e*, which shows that the signal is between 1,847 and 1,848 MHz. With the horizontal scale expanded to 100 kHz/cm, the signal frequency is read as 1,847.35 MHz (Fig. 16-38*f*). To obtain these displays, the output of the comb generator is superposed on the signal through a coaxial tee.

Modulation Measurement. With its frequency scan set to zero and the x axis representing time rather than frequency, the spectrum analyzer operates as a fixed tuned receiver to measure amplitude versus time. This has been called the *synchroscope mode*. When the analyzer is tuned to the carrier frequency with bandwidth at least twice that of the modulation frequency and with a linear display, the envelope of an AM signal can be observed as seen in Fig. 16-39. The modulation index m_a, where $m_a = 1$ for 100 percent modulation, can be calculated from the peak V_p and the trough V_T values as shown.

When the analyzer is operated normally, the two sidebands γ_s separated from the carrier, v_c at f_c, by the modulation frequency f_m are seen in Fig. 16-40. The modulation index can be calculated from the sideband and carrier amplitudes. The ability in the log mode to display signals differing as greatly as 70 dB makes it possible to measure AM

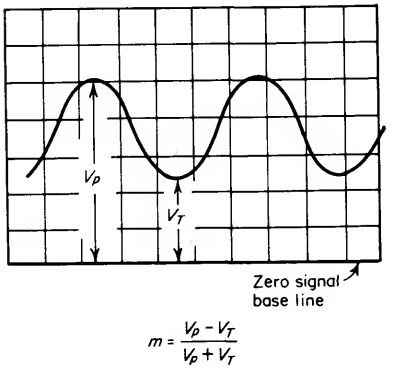


FIG 16-39 Envelope of AM seen on spectrum analyzer in zero-scan, or synchroscope, mode.

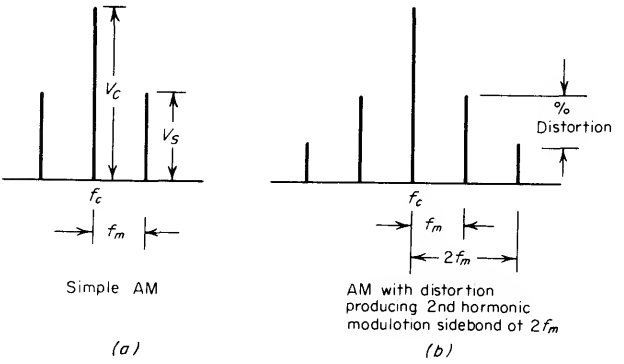


FIG 16-40 Spectrum of AM; f_c = carrier frequency, f_m = modulation frequency, and m_a = modulation index = $2V_s/V_c$.

as small as $m_a = 0.0006$. As seen in Fig. 16-40, it is also easy to measure distortion occurring in the modulation process.

A knowledge of the sideband configuration in FM enables one to calculate the FM index $\Delta\theta$ from the analyzer display. However, when $\Delta\theta$ (in radians) is less than about 0.2, the FM sidebands are identical with AM sidebands on the display, and the synchroscope mode may have to be used to make a distinction. For increased values of $\Delta\theta$, additional pairs of sidebands appear.

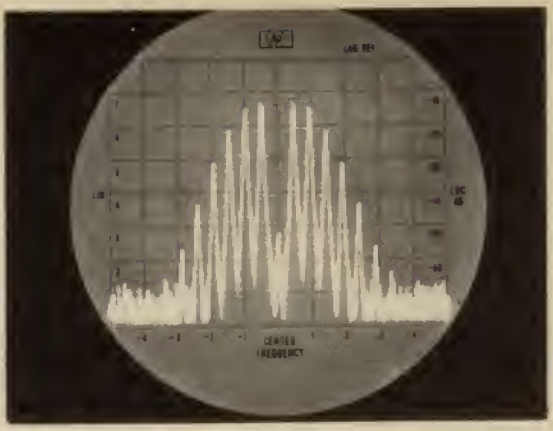
The amplitudes of both sidebands in a pair are equal when either pure AM or pure FM exists. Therefore, when inequality exists in the first pair, one can be sure that both FM and AM are occurring at the same modulation frequency. However, a phase and magnitude relationship *can* exist that results in sidebands of equal height, and so care must be exercised.

Figure 16-41 shows the spectra of two signals with FM. Modulation index can be expressed as

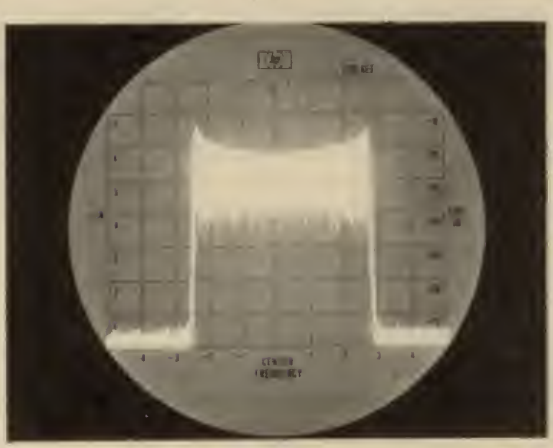
$$\Delta\theta = \frac{\Delta f_p}{f_m} \tag{16-13-1}$$

where Δf_p = peak frequency swing
f_m = modulation frequency

As modulation index is increased from a fractional value, carrier nulls occur at $\Delta\theta = 2.40$, 5.52, and $8.65 + n\pi$; so from the spectral display, $\Delta\theta$ is accurately known at these points, and by knowing f_m, the Δf peak can be accurately calculated, for instance in the calibration of an FM modulator. This is called the *carrier null method* of measuring frequency deviation.



(a)



(b)

FIG 16-41 Frequency-modulation displays on a spectrum analyzer: (*a*) Low-deviation FM. This is the spectrum for an FM signal at 10 MHz. The deviation has been adjusted for the second carrier null ($\Delta\phi = 5.6$). The sideband spacing is 25 kHz, and the modulation frequency, therefore, is $\Delta f_p = 5.6 \times 25$ kHz = 130 kHz (50 kHz per division at 10 MHz). (*b*) High-deviation FM. The transmission bandwidth required for this FM signal is 2.5 MHz (0.5 MHz per division). Expanding the scale reveals a sideband spacing of 10 kHz, the modulation frequency.

In most cases the peak-to-peak frequency deviation is approximately the width of the spectrum occupied by the FM signal. Usually the frequency deviation Δf peak is substantially greater than the modulation frequency f_m, or $\Delta\theta$ is substantially greater than 1. At very low modulation frequencies, subaudio, but with large relative deviation, the spectral display will be the signal sweeping slowly back and forth across the display by an amount exactly equal to twice the deviation. As the modulation frequency is now increased, this same display holds approximately true, as seen in the photograph of such a display in Fig. 16-41.

Continuous-wave Signal Frequency Stability and Spectral Purity. These measurements are measurements of unintentional, undesired FM. The spectrum analyzer and its local oscillators must have better frequency stability and spectral purity than the signal to be measured, and this should be determined beforehand from specifications or testing with a known stable signal.

The frequency drift of a signal is simply measured by observing the excursions of the signal across the display, as seen in Fig. 16-42. Over periods of minutes this measures long-term stability or drift; over periods of seconds, short-term stability or drift.

When the frequencies of undesired modulation-frequency components are greater than 1 Hz, so that the eye cannot follow them, but are lower than the resolution capabilities of the spectrum analyzer, the signal

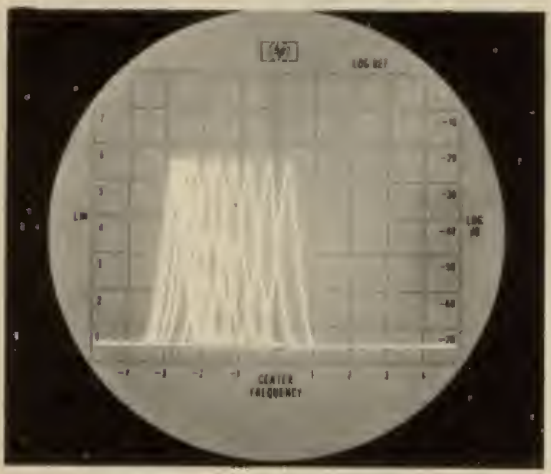


FIG 16-42 The frequency drift of an oscillator during warm-up is measured here. Scans are independently triggered 5 sec apart and are stored on the display section CRT. The frequency scale is 0.2 kHz per division with the center frequency of 20 MHz. The initial drift rate is 600 Hz/45 sec.

appears smeared as was seen in Fig. 16-25. This undesired FM coming from noise or power-supply ripple is sometimes referred to as *residual FM* and is measured as the width of the smear, which is the peak-to-peak frequency deviation.

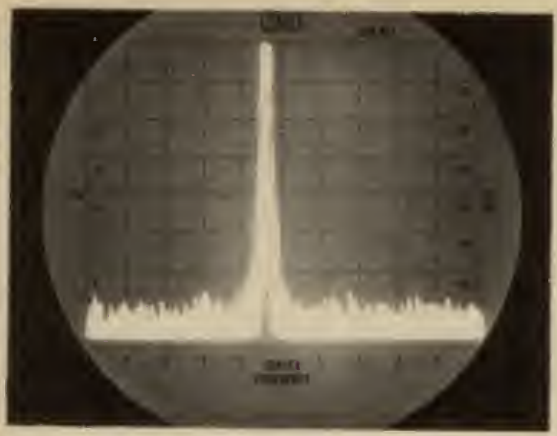
High-frequency wideband noise-modulation components produce what is sometimes called *phase noise* in a signal and result in noise sidebands continuously distributed on each side of the carrier, as seen in Figs. 16-25 and 16-43. Since these are distributed noise signals, they cannot be resolved as individual sidebands. Their voltage amplitude, but not that of the carrier, will be dependent on the spectrum-analyzer bandwidth used. The amplitude of these sidebands can be stated only in terms of their spectral density at various separations from the carrier. The information from this measurement can be given as the decibel level that the sideband peak envelope is down from the carrier decibel level at various frequency separations from the carrier, in which case the spectrum-analyzer bandwidth must be given. Or it can be given as the spectral power density in power per unit of bandwidth at various frequency separations from the carrier, in which case the carrier power level must be given.

Distortion and Noise. Since distortion in the transfer characteristic of a network affects the frequency components of a signal transmitted through that network, a spectrum analyzer can be used to make distortion measurements. The analyzer is often the most convenient instrument to use for this purpose. A full treatment of procedures will not be given, however, since distortion is treated in so many texts. The instruction manuals for commercial analyzers also give detailed instructions.

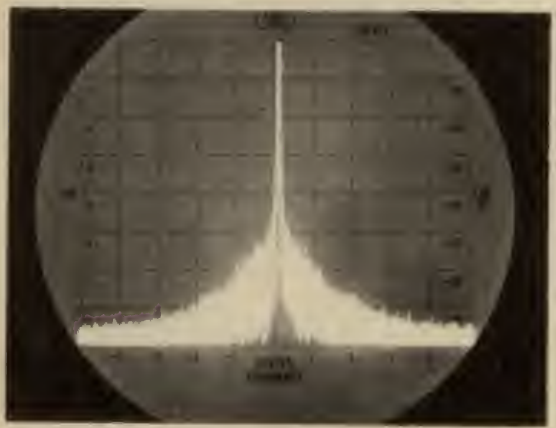
Similarly, white-noise measurement by means of an analyzer is fairly straightforward.

The noise bandwidth of the spectrum analyzer is determined, as shown in Fig. 16-44, by the usual method. A continuous-wave signal is applied to the input and the frequency axis expanded to display the analyzer selectivity characteristic. Either the "squared" mode of vertical display is used, or a linear voltage display is used and redrawn while squaring the ordinate.

Video filtering is filtering following the detector in a superheterodyne spectrum analyzer. Video filtering of a bandwidth much less than that of the intermediate-frequency filter is used when measuring white-noise density to average the display and make it readable at a definite value, which the normal grassy display of noise is not. If a squared mode of vertical display is used, the spectrum-analyzer detection is square law or power and no amplitude corrections need be made. The power is read from the vertical axis, directly if calibrated, or by continuous-wave substitution calibration if not, and divided by the noise bandwidth to obtain noise power density.

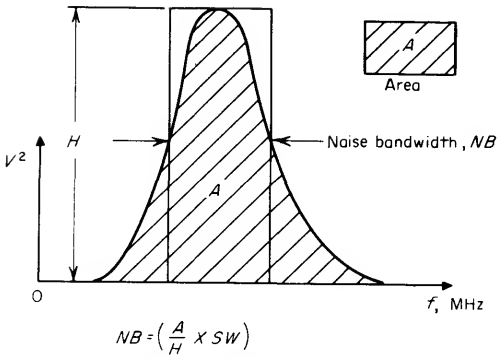


(a)



(b)

FIG 16-43 Here the spectral purity of oscillators is compared by a spectrum analyzer. In (*a*) the major noise sidebands are seen distributed around the carrier. The oscillator in (*b*) has a lower sideband noise level, but it has low-frequency residual FM. The sidebands are not resolved by the 0.3-kHz intermediate-frequency bandwidth; however, the deviation rate is low enough so that the back-and-forth movement of the continuous-wave signal is actually seen. The peak-to-peak frequency deviation is about 4 kHz (10 kHz per division), and since the center frequency is 100 MHz, stability is 4 parts in 10^5.



A = area of V^2 vs. f selectivity characteristic in cm^2
H = height, cm
SW = horizontal display calibration, kHz/cm

FIG 16-44 Noise bandwidth.

Caution must be observed that the spectrum analyzer input is not overloaded or even damaged when noise is being measured that extends over broad spectrum ranges. A density of 1 mW/MHz which extends over a range of 100 MHz is a total power of 100 mW, which could overload the instrument. The widest bandwidth consistent with resolution requirements should be used. Preselection helps by restricting spectral power entry into the spectrum analyzer. Overload can be found by the same method used with harmonic distortion. The display is valid if the insertion of x dB of input attenuation drops the entire display by x dB.

Measurement of Impulse Noise. Impulse noise is a broadband noise signal which in the time domain is a train of narrow pulses usually of low repetition rate and which may or may not be random in amplitude and rate. For instance, impulse noise is generated from voltage spikes created by engine ignition and electric motor commutation. As seen in Fig. 16-45, a train of impulses in the time domain produces a comb of equal-amplitude components extending over a wide range of the frequency domain. The components are spaced in the frequency domain by the pulse repetition frequency.

The duration of these pulses is usually short compared with the rise time of the systems with which, as noise, they interfere. Also, the pulses are spaced by time intervals that are long compared with the decay times of these systems. The response in these systems is thus an impulse response. The system responds to the area of the impulse and to each impulse individually. Impulse noise, as is white noise, is a signal distributed in frequency and so is characterized by its spectral density or

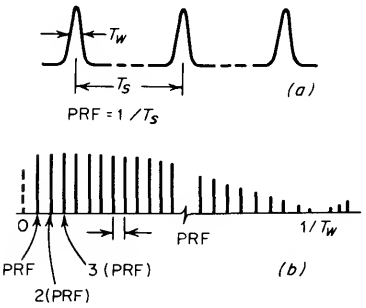


FIG 16-45 (*a*) Time-domain representation of pulse train. Pulse width is designated T_w, and pulse-rate period is T_s. (*b*) Frequency-domain representation of the pulse train shown above. With ideal pulses, frequency components would be a nearly identical amplitude out to $1/\pi T_w$ and may go through a null at a frequency of $1/T_w$.

spectral intensity. It is characterized by its voltage spectral intensity, which has units of volts per hertz or subunits thereof. The voltage referred to is the peak voltage that would be induced in a system with unit bandwidth by the impulse noise. With impulse noise, both the peak voltage and the average power increase linearly with bandwidth, while with white noise the voltage increases as the square root of the bandwidth and the average power increases linearly.

To measure impulse-noise spectral intensity, the spectrum-analyzer bandwidth must be wide compared with the pulse repetition rate, but narrow compared with the reciprocal of the pulse width. As seen in Fig. 16-46, the spectrum-analyzer bandwidth spans several spectral lines. In the figure,

$$IB = \frac{V_R}{SI} \tag{16-13-2}$$

where SI = spectral intensity
V_R = response of analyzer to impulse noise
IB = impulse bandwidth

Also in the figure,

A = time-domain area of impulse
V_A = amplitude of frequency-domain components of noise

The peak response of the spectrum analyzer to each pulse may be viewed as though at one instant all the components within its bandwidth added linearly to produce the peak response. So the spectral intensity is the number of lines per unit frequency multiplied by the amplitude of each line. The amplitudes of sinusoidal components given here are rms values. As seen in Fig. 16-46, the spectral intensity of an impulse in volts per hertz is $\sqrt{2}$ times the area of the impulse in volt-seconds.

Measurement of impulse noise is similar to measurement of white noise except that the impulse bandwidth IB is used; the video filter is not used because the peak response, not the average, is desired.

Impulse bandwidth can be determined in several ways. It can be arrived at graphically, as was shown for white-noise bandwidth in Fig. 16-40, but a linear, rather than squared voltage, axis is used. A better way is to apply a known spectral intensity, $\sqrt{2}$ times a known pulse area, and then divide the spectrum-analyzer response V_R by this known intensity to obtain the bandwidth. A third method is to apply a pulse signal of sufficient repetition rate that the spectrum analyzer set for a bandwidth less than the repetition frequency can measure the amplitude of the individual spectral components, V_A, as shown in Fig. 16-46. The spectral intensity of this signal is then the amplitude of one component divided by the repetition frequency. The analyzer is then set for the wide IB to be measured, and its response in voltage is divided by this known intensity to give the IB.

Even more precaution against overload should be observed than in measuring white noise, since peak voltages can be especially high and overload the analyzer input even though the distributed spectral intensity might be low. The widest bandwidth consistent with spectral-intensity envelope resolution requirements should be used. Overload is checked as always by seeing that changes in the display follow changes in input attenuation.

Pulse AM Signal Measurements. The spectrum of a carrier signal of frequency f_c modulated by a trapezoidal pulse train is shown in Fig. 16-47. The spectrum is centered at the carrier frequency f_c. The spectral components, lines, occur spaced from the carrier and from each other at multiples of the pulse repetition frequency prf. The shape of

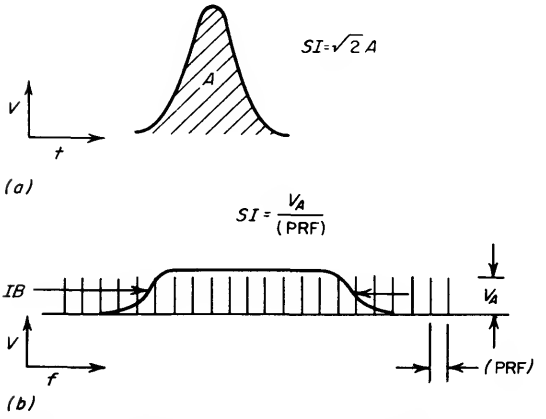


FIG 16-46 Spectral-analyzer impulse bandwidth.

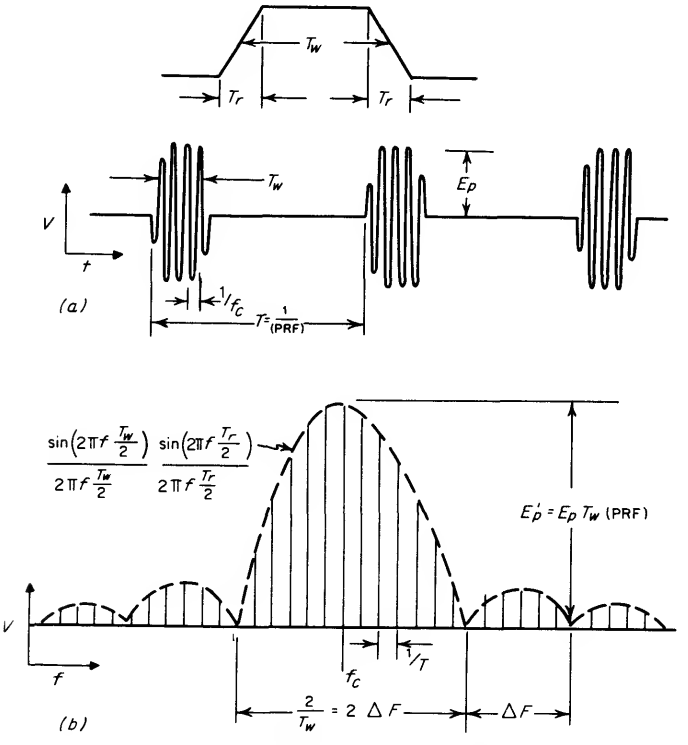


FIG 16-47 A sine wave modulated by trapezoidal pulses.

the spectrum, its envelope, is given by

$$\frac{\text{Sin } (2\pi f T_w/2)}{2\pi f T_w/2} \times \frac{\text{Sin } (2\pi f T_r/2)}{2\pi f T_r/2}$$

with nulls spaced at each side of the carrier frequency by multiples of the reciprocal of the pulse midwidth.

It is informative to examine, one by one, the effects that changes in the time-domain parameters of the signal have on the frequency-domain spectrum. If the carrier frequency is changed, the whole spectrum is completely unchanged in amplitude and frequency dimensions, but shifts to the new carrier frequency. If the pulse repetition frequency increases, the frequency dimensions of the envelope will be unchanged, but it will grow in amplitude in proportion to the pulse repetition frequency change and the spacing between spectral line components will increase and become equal to the new pulse repetition frequency. If the pulse mid-

width increases, the amplitude of the envelope increases in proportion and the frequency dimensions of the envelope shrink in proportion. If the shape of the trapezoid changes by changing T_r but not T_w, the shape of the spectrum envelope changes but not its amplitude at the center of the spectrum or the locations of the existing close-in nulls.

The responses that a modern swept superheterodyne spectrum analyzer can have to this periodically pulsed rf signal can be of two kinds and give two different but similar displays. One response is called a *line spectrum*, and the other is called a *pulse spectrum* or *spectral intensity plot* because it is the same kind of response as encountered with impulse noise described in the preceding section. Keep in mind that these are both responses to a periodically pulsed rf input signal and the *line* and *pulse* terminology refers to the response or display on the spectrum analyzer.

A line spectrum occurs when the spectrum-analyzer 3-dB bandwidth B is less than the most closely spaced spectral components of the input signal. Since the individual spectral components are spaced by the pulse repetition frequency of the periodically pulsed rf, this means that B of the spectrum analyzer must be substantially less than the pulse repetition frequency to get this type of display. All individual frequency components can be resolved; only one is within B at a time. The display is truly a frequency-domain display of the actual Fourier components of the input signal. Each component behaves as a continuous-wave signal would. The display has the normal true spectrum frequency-domain characteristics:

1. The spacing between lines on the display will not change when the analyzer sweep time display in centimeters per second is changed.
2. The amplitude of each line on the display, as measured by continuous-wave substitution, will not change as B is changed. (Of course, B must stay below the pulse repetition frequency to stay in the line spectrum mode.) The displayed height may change if the analyzer gain changes with B, but the measured signal amplitude will not.

A pulse spectrum occurs when B of the spectrum analyzer is greater than the pulse repetition frequency. The spectrum analyzer in this case cannot resolve actual individual Fourier frequency-domain components since several lines occur within its bandwidth, as with impulse noise. However, if B is narrow compared with the spectrum envelope, then the envelope can be resolved. The display is not a true frequency-domain display, but a combination of time and frequency display. It is a time-domain display for the display pulse lines since each pulse line occurs as each rf pulse occurs. The display lines occur at the actual pulse repetition frequency. It is a frequency-domain display of the spectrum envelope. The display has three distinguishing characteristics:

1. The spacing between the pulse lines on the display increases linearly

with the sweep speed. The pulse lines on the display occur at the pulse repetition frequency *PRF* and are spaced in real time by $1/PRF$. The shape of the spectrum envelope does not change with sweep speed.

2. The spacing between lines on the display does not change when the display width (in megahertz per centimeter) is changed. The spectrum envelope changes horizontally as one would expect.

3. Since the response is exactly as with impulse noise, the amplitude of the display envelope, as measured by continuous-wave substitution, increases linearly with a slope of 6 dB per octave as B is increased. This increase in amplitude with B persists until B equals about one-half the width of the main lobe of the spectrum envelope, and then no further increase occurs. At this point, B has collected almost all the spectral components. The display is now in the time domain. If the frequency scan were stopped by setting the spectrum width control to zero and one could sweep the display fast enough, one could begin to distinguish the time-domain shape of the pulse envelope. If B were increased further and became much wider than the main-lobe width, the detailed shape, in time domain, of the pulse envelope could be observed.

In the pulse spectrum just described, the response of the spectrum analyzer to each rf input pulse is the impulse response of the analyzer intermediate-frequency amplifier. The height of these impulse responses traces out the shape of the input spectrum as the front-end tuning is swept across the input spectrum. This impulse type of response is the explanation for the characteristics of this display.

Why use a pulse spectrum response? The spectrum envelope, its shape, amplitude, and spectral extent, is usually of interest instead of individual spectral component lines. The use of the pulse spectrum display with larger B gives a greater response than the line spectrum display does. The display amplitude at the center of the spectrum envelope with a line spectrum is, from Fig. 16-47, $T_w(\text{prf})E_p$ and with a pulse spectrum it is the product of the spectral intensity at the center of the spectrum and the impulse bandwidth IB of the analyzer, or $IB \times T_w \times E_p$, making

$$\frac{\text{Pulse spectral response}}{\text{Line spectral response}} = \frac{IB}{\text{prf}} \tag{16-13-3}$$

Therefore, the amplitude of the response increases linearly with B, as is always the case with an impulse response.

The analyzer's own noise voltage level will increase but as $\sqrt{B}$, so the increase in signal-to-noise ratio increases similarly. Operating with B greater than the pulse repetition frequency increases sensitivity to pulsed rf signals and increases the dynamic range. The input level should be

held to a value such that the peak pulse signal barely does not overload the input mixer.

16-14 Some Rules of Thumb for Choosing Bandwidth and Other Control Settings when Viewing Pulsed RF Spectrums

Sensitivity versus Envelope Resolution. The B should be as wide as possible for the highest sensitivity but not wider than $0.1/T_w$, the pulse midwidth, or the resolution of the spectrum envelope will be seriously impaired. Another way of saying this is that B should be less than 5 percent of the width of the spectrum-envelope main lobe. If higher resolution is required to look at faster falling spectrum envelope skirts or to look more deeply into nulls, narrower B is required. To resolve 20 to 30 dB into nulls, a good rule is

$$B < \frac{0.03}{T_w} \tag{16-14-1}$$

Avoidance of Base-line Lifting. This phenomenon, shown in Fig. 16-48, occurs with a pulse spectrum when the response from one pulse has not fully decayed in the spectrum analyzer when the next pulse occurs. As mentioned before, the intermediate-frequency amplifier decay-time constant is $0.3/B$, and for the pulse to decay down to 1 percent requires five time constants. To meet this requirement, the rule is

$$B > 1.7 \text{ prf} \tag{16-14-2}$$

This will keep the lifted base line at least 40 dB below the spectrum envelope.

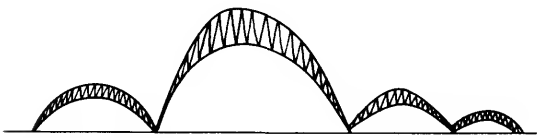


FIG 16-48 Base-line lifting.

Avoidance of Envelope Peak Fluctuation. Fluctuation of the spectrum envelope peak on the display occurs when the pulse does not occur during the time the analyzer is sweeping through the peak of the spectrum envelope on each sweep. This occurs with fast sweeps, low pulse repetition frequency, and wide pulses, and the fluctuation looks almost random at times and as if something is cutting on and off. To avoid this, the time to scan through 10 percent of the main-lobe width must be greater than

1/prf. The rule is

$$\frac{df}{dt} < 0.2\text{prf}\ \Delta F \tag{16-14-3}$$

$$\frac{df}{dt} < \frac{0.2\text{prf}}{T_w} \tag{16-14-4}$$

where $\frac{df}{dt}$ = analyzer-frequency sweep velocity, Hz/sec

ΔF = minor-lobe null separation, Hz; the main-lobe base width $= 2\Delta F$

T_w = rf pulse width, sec

This rule assures that the peak will be displayed on each sweep and avoids having to watch several sweeps to catch the peak value of the displayed spectrum.

Loss in Sensitivity Due to Sweeping Too Fast. If one sweeps a spectrum analyzer past a continuous-wave signal too fast, a loss in response occurs. The rule for avoiding this sweep desensitization, already explained, is

$$\frac{df}{dt} < B^2 \tag{16-14-5}$$

Loss in sensitivity can also occur from sweeping too fast through the pulse spectrum of a periodically pulsed rf signal. The rule for less than 1 dB of loss in sensitivity is that

$$\frac{df}{dt} < \frac{2.5}{T_w^2} \tag{16-14-6}$$

Frequency Domain. Most of the preceding has dealt with the pulse-spectrum mode of response. The area between it and the line-spectrum mode of response is a gray area that does not have quite the characteristics of either type of display. In this area, interpretation of the display is difficult and is best avoided.

The criterion for being in the line-spectrum response mode when viewing a periodically pulsed rf signal is that

$$B < \frac{\text{prf}}{10} \tag{16-14-7}$$

It is required in this case that there be no decay in response between pulses, and the above allows less than 3 dB of decay.

The criterion for being in the pulse-spectrum mode is that $B >$ prf, so there is a decade of gray area to avoid.

Overload must be again particularly guarded against with the pulsed rf inputs described to avoid having the pulse peak overload the spectrum analyzer input.

16-15 Electromagnetic Interference Measurements

Electromagnetic interference measurements are spectrum analysis measurements of both continuous-wave and broadband or impulse noise signals radiated from or conducted by cables from a piece of equipment. The measurements are very similar to those already described, but antennas and current transformer probes are used as transducers ahead of the spectrum analyzer, which enables it to measure field strength and current. Table 16-1 tabulates some of the types of antennas used. The spectrum-analyzer amplitude readings are converted to the field strength at the antenna through the use of the antenna factor K, where

$$K = \frac{\text{electric field strength at antenna, V/m}}{\text{signal delivered by antenna to instrument, V}}$$

Current measurements are made by using an rf current transformer.

TABLE 16-1 Antennas Specified in MIL-STD-461 for Emission and Susceptibility Measurements

Frequency	*Description*
14 kHz–25 MHz	41-in. rod (electrical length ½ m) and matching network.
20–200 MHz	Biconical antenna
200–1,000 MHz	Conical logarithmic spiral antenna
1–10 GHz	Conical logarithmic spiral antenna

CHAPTER SEVENTEEN

MICROWAVE NETWORK ANALYSIS

From notes by

Stephen F. Adam

Douglas K. Rytting

William Heinz

Hewlett-Packard Company
Palo Alto, California

Microwave devices and networks are characterized by impedances and gains or attenuations, just as devices and networks are characterized at lower frequencies. However, both the methods of measurement and the parameters measured are profoundly affected by the fact that a wavelength at microwave frequency is short in comparison with the dimensions of the system. It becomes imperative to consider the existence of waves traveling at finite velocities and being partially transmitted, partially reflected, at impedance discontinuities in a system.

Microwave impedance and transmission measurements are also conditioned by the universal practice of transmitting microwave signals through virtually lossless, uniform transmission lines having real, positive

characteristic impedances Z_0 or R_0. For years, instruments have been designed to measure the reflection coefficient Γ at a reflection plane or discontinuity in such a transmission system, where $\Gamma = E_r/E_i$. The incident voltage E_i and reflected voltage E_r are phasors in general, as in Γ, but the absolute value of reflection coefficient ρ is often useful. It is simply $|\Gamma|$ at the reflection plane.

Microwave transmission systems are also characterized by standing waves along lines, since E_i and E_r are voltages traveling in opposite directions. Standing-wave measurements have also been useful for many years.

In this chapter it is assumed that the reader understands classical transmission line theory and has some acquaintance with simple devices such as slotted lines, tunable and fixed loads, and detectors. After a brief review of basic techniques, we shall proceed to a treatment of more highly developed microwave devices used to measure transmission parameters. Particular emphasis is placed upon the use of the s parameters in measurement and design. Last, the trend toward automatic and computerized network analysis is viewed.

17-1 Reflection and Impedance Measurement

The measurement easiest to understand physically is the voltage–standing-wave ratio, vswr. One way to measure vswr is to use a slotted line, a transmission line or section of waveguide with provision for sensing the voltage on a small sliding probe that projects into the line without introducing significant reflections of its own. The other approach is to measure the amplitudes of incident and reflected waves separately by placing two directional couplers in the transmission system, back to back. We assume that the devices and networks being tested are properly terminated, but the importance and complexity of correct termination should not be overlooked, especially in broadband systems.

Slotted Lines. A slotted line [1] can be defined as a substantially lossless, uniform precision section of transmission line provided with a longitudinal aperture allowing a signal pickup probe to penetrate into the fields on the line, which enables one to detect the field strength variations along that transmission line.

Coaxial slotted lines, using true coaxial cross section, are severely limited in their performance because of the finite width of the slot to enable the probe to penetrate into the line. This causes slight deviations of characteristic impedance and excess leakage through the slot. Wholey and Eldred [2] suggested a slab line arrangement, now widely used, a cross section of which is shown in Fig. 17-1. With proper design, Z_0 can be maintained even when a sliding probe is used.

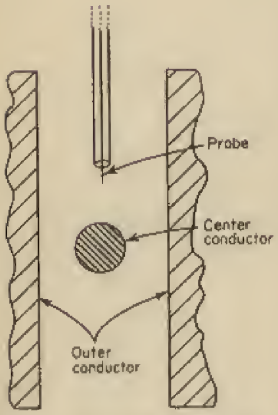


FIG 17-1 Cross section of a slab transmission line.

FIG 17-2 Slotted-line detector probe, Hewlett-Packard Company 447A.

A swept-frequency [3, 4] vswr technique has been developed in the past few years. Figure 17-3 shows the block diagram of such a measurement setup. The logarithmic display provides information of the vswr in decibels. Slotted lines generally probe the electric field distribution. Modern coaxial slotted lines use broadband, untuned pickup probes, such as shown in Fig. 17-2.

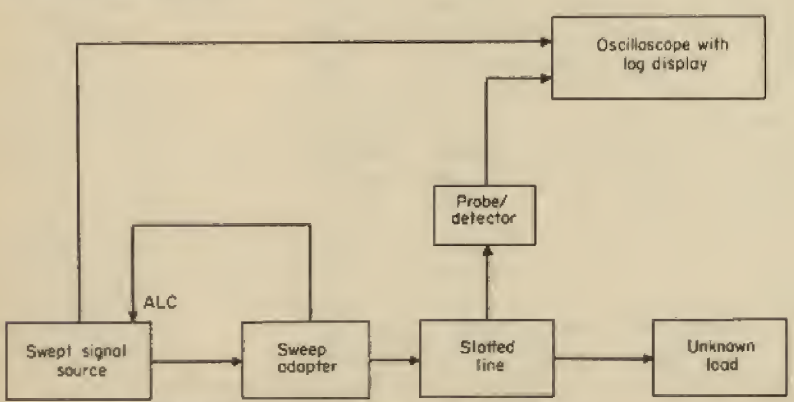


FIG 17-3 Block diagram for measuring SWR with swept-frequency slotted-line technique.

Standing-wave indicators are available directly calibrated in vswr, taking into account the square law of the crystal detector used normally in a slotted-line probe. These standing-wave indicators usually consist of a precision step attenuator and a narrow-band 1 kHz tuned amplifier capable of amplifying very small signals.

When measuring vswr on a slotted line, the ratio of the maximum to the minimum value of the voltage standing-wave pattern is determined. Errors in such measurements generally can be defined in terms of the reflection coefficient as follows [5]:

$$\pm\Delta\rho = A + B\rho + C\rho^2 \qquad (17\text{-}1\text{-}1)$$

where A = slope and loss of slotted line and residual reflections in the line
B = probe reflection, attenuator accuracy, and square-law error of detector
C = probe reflection

Determining the actual shift of the minimum of the standing-wave pattern compared with a reference plane on the slotted line provides information about the phase of the impedance of the termination. Data gathered from such measurements can be plotted on a Smith chart. The frequency is swept rapidly back and forth, and a storage or a variable-persistence cathode-ray oscilloscope is used. The carriage of the slotted line is moved a distance of at least one-half of a wavelength at the lowest frequency of interest. This makes the vertical height of the envelope of the display pattern represent the vswr in decibels versus frequency.

Reflectometers. As mentioned above, another way to study impedance discontinuity in a line is to separate the reflected component of voltage from the incident component. The basic device for separation is the directional coupler, which gives a signal in its auxiliary arm proportional to the signal component traveling in only one direction in the main line. Signals coming from the opposite direction produce no output in the auxiliary arm (or very little). Two directional couplers are connected back to back, as shown in Fig. 17-4. The outputs, after detection, are fed to a ratiometer, which indicates the reflection coefficient. Remember

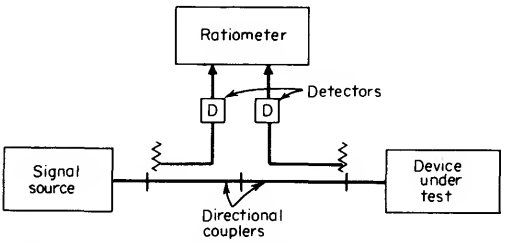


FIG 17-4 Reflectometer arrangement.

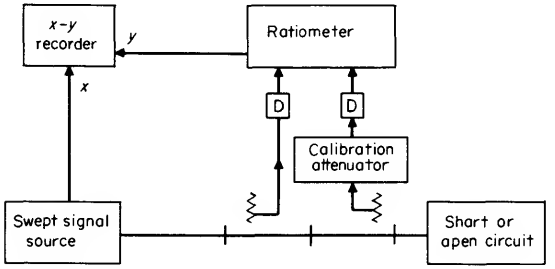


FIG 17-5 Calibrated reflectometer.

that the relationship between vswr, or σ, and reflection coefficient ρ is

$$\sigma = \frac{1+\rho}{1-\rho} \quad \text{or} \quad \rho = \frac{\sigma - 1}{\sigma + 1} \tag{17-1-2}$$

To make the reflectometer of Fig. 17-4 accurate over a broad frequency band imposes some strict requirements upon the directional couplers. If the coupling ratio (ratio of signal level in auxiliary arm to level in forward direction) varies with frequency, it must vary almost identically in both couplers. Further, high directivity is required, and the couplers must cause negligible mismatch of impedance in the network.

In order to allow some tracking error, calibration procedures have been developed to produce calibration grids by using short and open circuits at the measurement port. To produce calibration lines, attenuation values of equivalent return losses are inserted into the auxiliary arm of the coupler providing the reflected voltage. Such a setup is shown in Fig. 17-5.

With the advent of leveled signal sources the use of ratiometers has been largely eliminated. Leveled reflectometers do not need to make the

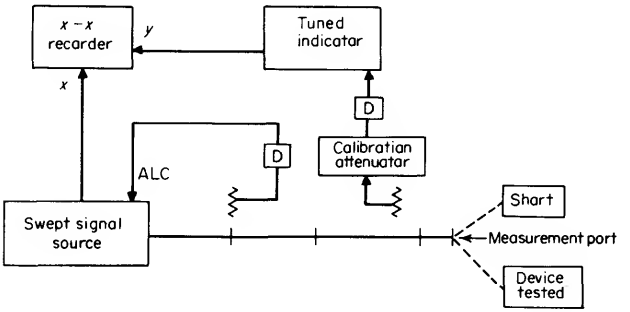


FIG 17-6 Leveled reflectometer.

TABLE 17-1 Error Factors for Reflection Measurements

Error factor	Calibration grid	No calibration grid
A.....	Directivity; attenuator accuracy	Directivity
B.....	Directivity; attenuator accuracy	Directivity; attenuator accuracy Coupler/detector tracking; detector square-law error
C.....	Reflectometer source match	Reflectometer source match

actual ratio measurement, since calibration grids can be established by using shorts and open circuits to calibrate. Such a setup is shown in Fig. 17-6. To find the errors of such measurement the same basic error can be used as in the slotted-line case [6]. That is,

$$\Delta\rho = A + B\rho + C\rho^2 \tag{17-1-3}$$

where ρ is the measured magnitude of reflection coefficient.

There are two basic techniques, depending upon whether a calibration grid is used. Therefore, the error factors are as in Table 17-1.

17-2 Attenuation Measurements

Measurement of attenuation or gain can be divided into three basic classes: rf substitution, audio or dc substitution (square-law detection), and intermediate-frequency substitution (linear detection). Beatty [7] defines attenuation as the decrease in power level (at the load) caused by inserting a device between a Z_0 source and Z_0 load.

The rf substitution technique uses a standard known microwave attenuator for comparison (Fig. 17-7).

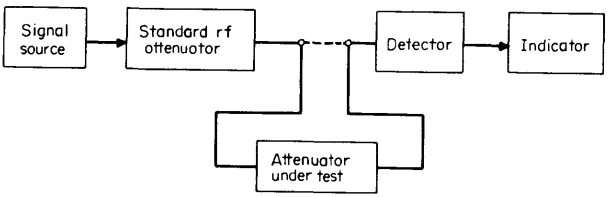


FIG 17-7 Attenuation measurement with rf substitution.

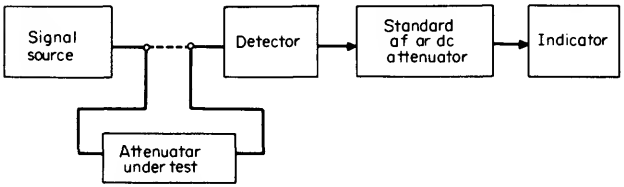
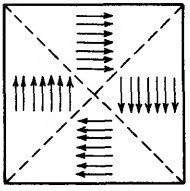


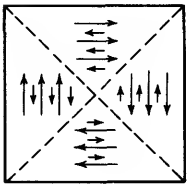
FIG 17-8 Audio frequency or dc substitution.

Audio-frequency or dc substitution uses low-frequency or dc attenuators to compare microwave attenuation (Fig. 17-8). This technique takes advantage of more accurately calibrated attenuators. Of course, in this method the error in the detector law does come into consideration, and great care has to be taken to keep detectors in their square-law region, while in the rf substitution method the detector always operates at the same level, which eliminates the need to worry about detector law. The combination of the two techniques is often advantageous.

The techniques above are quite limited in their dynamic range, their sensitivity reaching only to −40 to −60 dBm. Instruments with linear detection increase the sensitivity to −100 dBm or further and therefore greatly expand the dynamic range.



(a)



(b)

FIG 17-9 Possible single crystal domain configuration showing alignment for (a) ferromagnetic and (b) ferrimagnetic cases (no applied field).

Errors in attenuation measurements can be classified as scalar, vectorial, or random (uncertainty).

Scalar errors are the ones that could be calibrated out, if known. For example, the error of the standard attenuator used is a scalar term, as is drift of the signal source and the detector, detector law, etc.

Vectorial errors are the type that do vary with frequency and can be represented by rotating vectors. Such errors are multiple mismatch errors. With varying frequency and varying electrical length, they beat with each other. If their values are known in both magnitude and phase at each frequency, they can be calibrated out by vectorial manipulations. These errors, if they are not taken closely into consideration, can be quite large, and most of the time they cause the largest errors in such measurements.

Uncertainty terms or random errors are those that cannot be predetermined. Such errors are caused by variations—connector repeatability, instrument readability, resettability, etc. Many repeated mea-

urements and the averaging of the results can cut down the size of these errors. There are statistical methods taking these into account.

17-3 Ferrite Devices

Directional couplers, discussed above, yield output signals with amplitudes depending upon the direction of propagation in the main channels, but these couplers are passive, static devices in which reciprocity holds. There is also an important class of microwave devices based on materials that have transmission characteristics depending upon the direction of travel of waves in them. These materials are *ferrites*. The technology of nonreciprocal microwave devices is only about 20 years old. Prior to this period, reciprocity was considered a fundamental property of all linear passive networks. Nonreciprocity permits differential phase shifting, attenuation, and channeling of waves depending on their direction of propagation through the device.

Linear ferrite devices are usually of three types: isolators, circulators, and phase shifters. These will be briefly described below, after a basic description of the properties of ferrites.

Composition and Physical Properties of Ferrites. Ferrites are magnetic oxides of iron, commonly containing other metals or rare earths, or both, to achieve any of several unique properties. They are usually prepared from oxides of their constituent elements in finely powdered form, and are mixed, pressed to shape, and then sintered [8]. The result is a hard, brittle, ceramiclike material combining magnetic properties and high electrical resistance.

Three basic crystal structures are most common:

Spinels. These are of the general form MFe_2O_{12}, where M is a divalent metal ion. The crystals have a cubic symmetry. Spinels are generally used at low frequencies and at microwave frequencies.

Garnets. The general form for the garnets is $X_3Fe_5O_4$, where X is an ion of yttrium or another rare earth. Yttrium-iron-garnet (yig) has found widespread use at microwave frequencies. It can be doped with other rare earths or metals to produce a variety of properties.

Hexagonals. When the M ion described above has a diameter much larger than the Fe atoms (such as with barium), the crystal has a hexagonal structure. Barium ferrites are used as permanent magnet materials.

Ferromagnetism and Ferrimagnetism. The *magnetic* behavior of ferrites, just as of any other magnetic material, is fundamentally due to the property of the spinning electron. The relationships between the individual atoms determine the macroscopic behavior of the material.

In ferromagnetic materials, the individual magnetic moments align themselves in parallel within the crystal and form domains (see

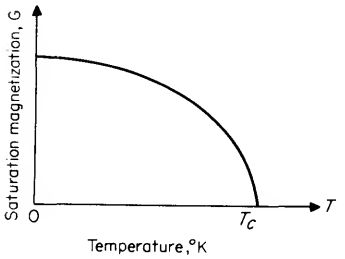


FIG 17-10 Saturation magnetization versus temperature.

Fig. 17-9*a*). These configurations conform to a minimum total energy in the sample. Increasing temperature causes less and less alignment within each domain, owing to thermal vibrations of atoms in the lattice, until the Curie temperature T_c is reached, where the magnetization becomes zero.

Ferrimagnetic materials (ferrites) are different from ferromagnetic materials in that the spins align themselves antiparallel to each other (see Fig. 17-9*b*). A net magnetization is produced if the oppositely aligned moments are different in magnitude. Just as in the ferromagnetic case, the material will be divided into domains which are usually randomly oriented when no external field is applied. As an increasing field is applied, the domains will reorientate themselves more and more in the direction of the field, until they are all aligned and the material is "saturated." The bulk sample may then be characterized macroscopically by a saturation magnetization $4\pi M_s$, which has the units of gauss in the commonly used gaussian system of units. Figure 17-10 shows how $4\pi M_s$ varies with temperature.

Ferrimagnetic Resonance. When a dc magnetic field is applied to a ferrite, the individual moments align themselves in the field. In doing so, however, they precess about the field as a spinning top precesses in the earth's gravitational field when disturbed (see Fig. 17-11). The frequency of precession ω_0 is directly proportional to the magnitude of the

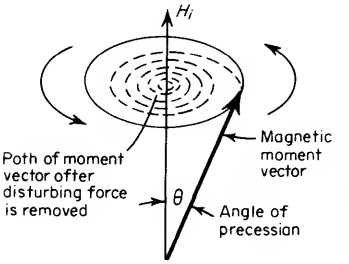


FIG 17-11 Precessing magnetic moment vector.

field inside the ferrite, H_i or

$$\omega_0 = \gamma H_i \tag{17-3-1}$$

where $\gamma = 2.8$ MHz/Oe. Of course the motion is damped and the precessing magnetization vector spirals in and eventually becomes aligned with the applied field. The damping represents transfer of energy from the precessing electrons to the lattice, where it is dissipated as heat.

If microwave energy is applied in such a manner that a circularly polarized (CP) component of rf magnetic field exists in the same direction as the natural precession, energy will be delivered to the precessing spins, and to the lattice by the damping mechanism. Maximum energy absorption occurs when the frequency of the wave equals the natural frequency of precession as determined by Eq. (17-3-1). No energy is absorbed from an rf field circularly polarized in the opposite sense.

Since the rf permeability is a function of the direction of the rf magnetic field relative to the applied dc field, it is in general a tensor quantity. If we restrict our attention to the case of CP rf magnetic fields in a plane perpendicular to the dc field, the permeabilities are scalars,

$$b_\pm = \mu_\pm h_\pm \tag{17-3-2}$$

where $\mu_\pm = \mu'_\pm - j\mu''_\pm$, and where the subscripts denote the two senses of polarization. The permeabilities are plotted as functions of dc magnetic field in Fig. 17-12. The term μ''_+, the imaginary part of μ_+, represents the

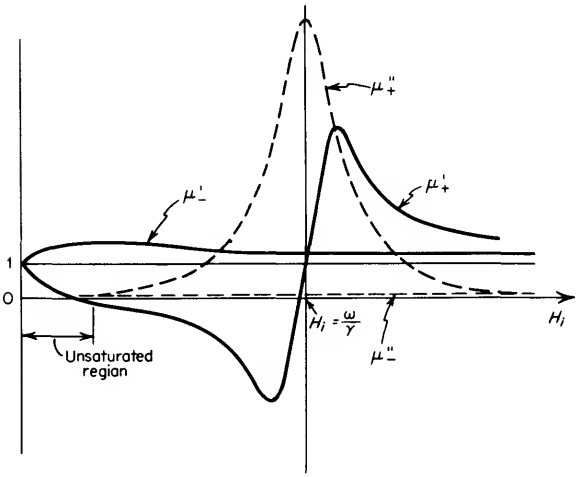


FIG 17-12 Real and imaginary components of μ_+ and μ_- versus applied magnetic field.

lossy component of the permeability and is thus maximum when the excitation frequency equals the resonance frequency ω_0. The width of the μ_+'' curve is a measure of the line width ΔH of the material. This is directly proportional to the amount of damping of the precessing spins.

The nonreciprocal devices to be described below make use of the difference in both the real and imaginary parts of μ_+ and μ_-.

Low-field Losses. Microwave ferrite devices can incorporate ferrite materials biased to ferrimagnetic resonance (making use of the difference in absorption between left- and right-hand-circulating waves), or they can be operated at fields above or below resonance (making use of the difference in the real part of the permeabilities). An important limitation exists on the low-frequency operation of these devices and must be mentioned at this point. This is the phenomenon of low-field losses.

Because of the domain structure of ferrites having applied fields below the value required for saturation and because of the presence of internal fields, resonance losses can occur at low microwave frequencies. These losses will be independent of the sense of polarization of the microwave magnetic fields and will thus degrade the performance of the device [9]. Low-field losses are most troublesome for devices operating below resonance, since they increase with decreasing dc field when the frequency of operation is $\omega < \omega_{LF}$, where

$$\omega_{LF} = \gamma[H_a + 4\pi M_s] \tag{17-3-3}$$

In Eq. (17-3-3), γ is the gyromagnetic ratio, $4\pi M_s$ is the saturation magnetization, and H_a is the internal anisotropy field. Typically H_a is of the order of 100 Oe or less for microwave ferrites.

Biasing above resonance will permit operation below this frequency if the ferrite can be saturated at the higher magnetic field. However, certain devices cannot be operated in this mode without suffering degradation in other parameters, such as bandwidth.

Isolators. An isolator is a two-port device that allows the propagation of energy in one direction with low loss and has high attenuation for power passing in the other direction. It can easily be shown from the unitary property of the scattering matrix that this function cannot be performed by a lossless two-port network [10]. Thus, the reverse attenuation must be in the form of dissipative loss.

Isolators are useful as stabilizers and protectors of microwave sources. Reflections from mismatched loads can affect the frequency and phase of an oscillator. With high-power tubes, excessive reflected power can result in damage or reduced tube life. The reverse attenuation of an isolator will reduce the level of power reflected back to the source and thus will present a better, more uniform match to the source. Of course, this

function can be performed by an attenuator (pad), but this is done at the expense of reduced power available to the load.

There are several common types of isolators. Resonance isolators contain ferrite biased to ferrimagnetic resonance in microwave transmission structures providing regions of circular polarization (CP) of magnetic field that have opposite senses for the two directions of propagation. Power is absorbed by the ferrite when the CP is in the same sense as the precession, and little interaction occurs for CP in the opposite sense.

A rectangular waveguide supporting a wave propagating in the dominant H_{10} mode has two regions of CP, one right-handed and one left-handed. For a wave traveling in the opposite direction, the senses of CP at these points are reversed. Figure 17-13*b* shows a simple isolator containing two ferrite slabs. Resonance isolators may use from one to four such slabs located against the broad walls on one side of the waveguide. When slabs are also located on the left side of the guide, the magnetic field applied to them must be in the opposite direction since the direction of CP is also opposite from that on the right. Dielectric loading, such as is shown in Fig. 17-13*c*, has been found to improve reverse attenuation and bandwidth [11, 12].

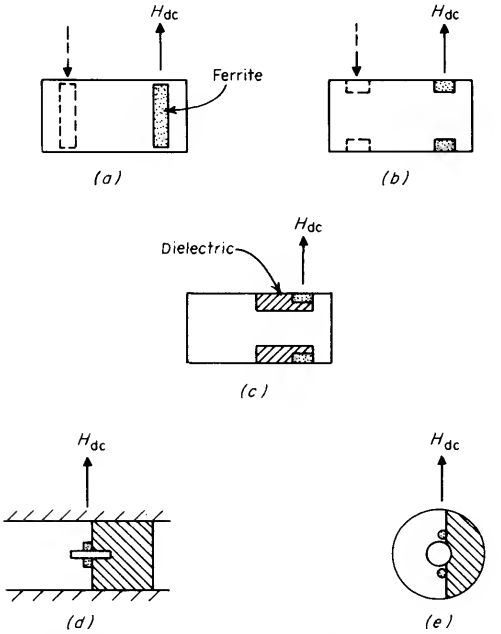


FIG 17-13 Resonance isolator configurations.

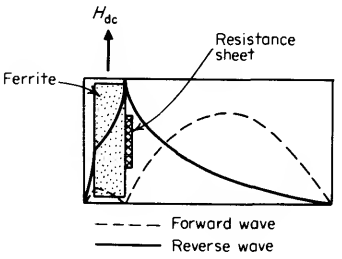
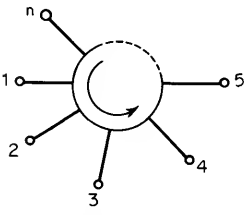


FIG 17-14 Field displacement isolator configuration showing electric field patterns for forward and reverse waves.

Figure 17-13*d* and *e* shows devices in which TEM structures are loaded asymmetrically with dielectric to produce H modes having regions of CP [13, 14].

Figure 17-14 shows a ferrite-loaded waveguide and the electric field patterns for forward and reverse propagation. In order to achieve these patterns, the ferrite is biased below resonance, and use is made of the difference between the real parts of μ_+ and μ_-. The resistive sheet absorbs power from the reverse wave since the electric field is large, but very little forward attenuation occurs since the forward electric field is nearly zero.

The main advantages of this type of isolator over the resonance isolator are extremely low insertion loss and a lower applied magnetic field [15, 16]. Insertion losses as low as 0.1 dB over bandwidths in excess of 10 percent have been obtained.



(a)

$$S = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & & & \cdots & 0 \end{bmatrix}$$

(b)

FIG 17-15 (*a*) **The** n**-port circulator symbol for 1 → 2 → 3 → · · · →** n **circulation;** (*b*) **scattering matrix of** n**-port circulator shown above.**

Circulators. An ideal circulator is a device having three or more matched ports in which power incident on any port is conducted out of the next port in one particular direction with no loss and with no coupling to any other port. An n-port circulator is shown symbolically in Fig. 17-15*a*, in which the ports are numbered so that power incident on port 1 flows out of port 2, power into port 2 flows out of port 3, etc., with power into port n emerging from port 1. The scattering matrix of the circulator is shown in Fig. 17-15*b*, along with the defining relationships from which it is obtained [17].

The simplest type of circulator both conceptually and practically is the three-port symmetrical Y-junction circulator. Three versions of this device are shown in Fig. 17-16.

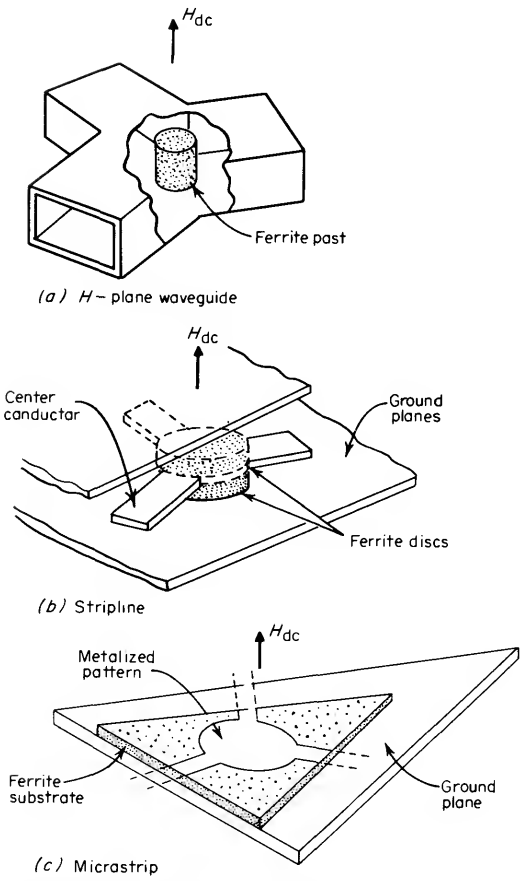


FIG 17-16 The Y-junction circulator. (*a*) H-plane waveguide, (*b*) stripline, (*c*) microstrip.

These devices all depend on a resonating junction containing ferrite magnetized along the axis of symmetry. At microwave frequencies, the magnetic field applied is usually below resonance, while above-resonance operation is used at ultrahigh frequencies to avoid low-field losses.

The currently accepted theory of operation of these devices [18, 19] involves splitting of modes in the junction due to the gyrotropic ferrite. This is illustrated in Fig. 17-17. In Fig. 17-17*a*, the field configuration of the standing-wave pattern of the lowest-frequency mode in a cylindrical resonator is shown. The E field, represented by crosses and dots, is

perpendicular to the plane of the disk, and the H field, represented by lines, is entirely in the disk plane.

There is no variation of fields along the axis of the disk. Under suitable conditions, application of a magnetic field causes the pattern to be rotated 30° as shown in *b*, where the E field at port 3 is zero. In this case, energy is transmitted from port 1 to port 2, and port 3 is isolated. The standing-wave pattern in Fig. 17-17*a* can be generated by two counterrotating patterns (modes), in which there is CP of the magnetic field in the center of the disk. When the dc magnetic field is applied, the different permeabilities for the two modes cause their resonant frequencies to split apart. The proper amount of splitting (which is related to the Q of the resonator) will cause the desired 30° shift in the standing-wave pattern, as shown in Fig. 17-17*b*. The analysis of wideband circulators requires consideration of higher-order modes as well.

To date, optimum performance at microwave frequencies has been achieved by using the strip-line structure [20]. Strip-line devices having isolation in excess of 20 dB over octave bandwidths are available, with insertion loss of a few tenths of a decibel. Lumped constant circulators, employing ferrites biased above resonance, have been built to operate at ultrahigh frequencies [21] down to frequencies below 50 MHz.

Some of the applications of three-port circulators are illustrated in Fig. 17-18. Figure 17-18*a* shows how connection of a termination to one port makes the device an isolator. Power into port 2 is diverted to port

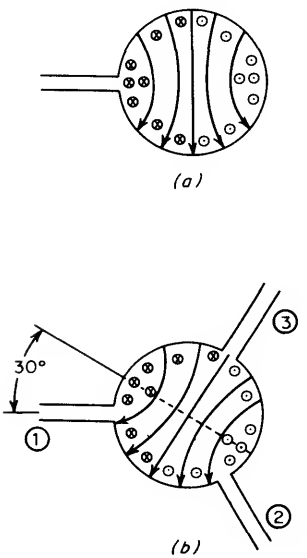


FIG 17-17 (*a*) **Standing-wave pattern, no field applied;** (*b*) **standing wave pattern shifted 30° upon application of magnetic field.**

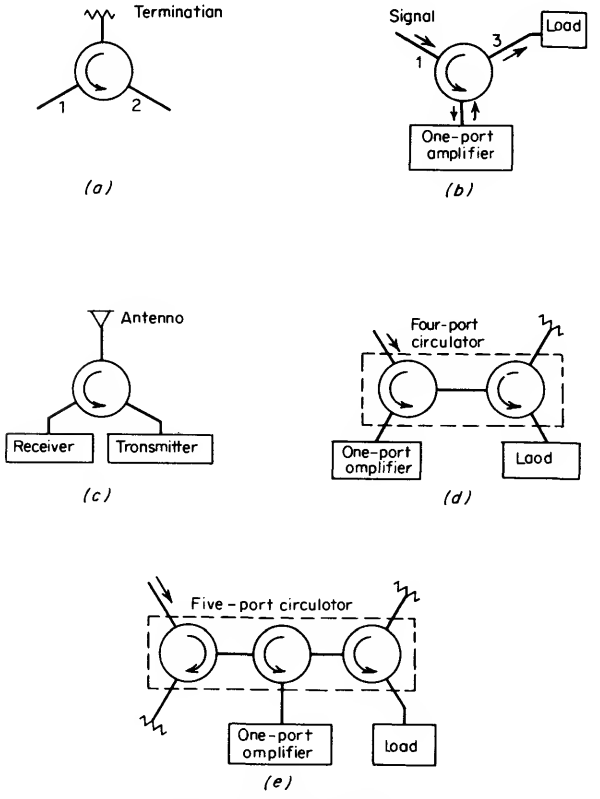


FIG 17-18 Three-port circulator applications.

3 and dissipated in the load. The reflection coefficient of the load must be low, since the return loss is directly equal to the isolation. Figure 17-18*b* shows a one-port reflection amplifier (such as a parametric amplifier or cavity maser) connected to a circulator. The wave reflected from the amplifier is the amplified output and goes to port 3 and the load. Figure 17-18*c* shows a transmitter-receiver system using a common antenna. Isolation of the receiver is however limited by the vswr of the antenna. Figure 17-18*d* and *e* shows how three-port circulators may be used to realize a four-port and five-port circulator respectively, and how they are typically used to achieve isolation of the load or source, or both.

Ferrite Phase Shifters. A gyrator is a two-port device having 180° more phase shift of a wave traveling in one direction than of a wave traveling in the opposite direction. The gyrator circuit symbol is shown in Fig.

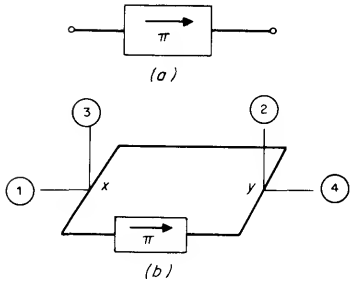


FIG 17-19 (*a*) A gyrator circuit symbol. (*b*) Four-port circulator realized from two magic T's and a gyrator. Phase shift from y to x is assumed equal in each leg.

17-19*a*, and part *b* of the figure shows how a four-port circulator may be realized from a gyrator and two magic T junctions. Other combinations of hybrid junctions and other values of differential phase shift may be used in the synthesis of four-port circulators [10, 21], but these will not be discussed here.

If the magnetic field applied to the waveguide device in Fig. 17-13*a* is below the resonance value, the different dispersive (real) parts of μ_+ and μ_- will yield two different propagation constants β_+ and β_- for the two directions of propagation. Thus the differential phase shift $\Delta\phi = (\beta_+ - \beta_-)l$, where l is the active length of the device.

The device shown in Fig. 17-20 is a digital phase shifter capable of being switched rapidly between two states. When a single current pulse is applied to the wire, the ferrite is magnetized in one direction and remains in its remanent state. Since the magnetization will be as shown by the solid arrow, the conditions for differential phase shift are satisfied [22]. Application of a current pulse in the opposite direction reverses the magnetization and, thus, the differential phase shift. Such devices are useful for phased array antennas.

The device shown in Fig. 17-21, in which the rod is magnetized axially, will yield reciprocal phase shift which varies with applied magnetic field [23]. It is symmetrically located in the waveguide, and the magnetization is uniform in the ferrite. This device is reciprocal.

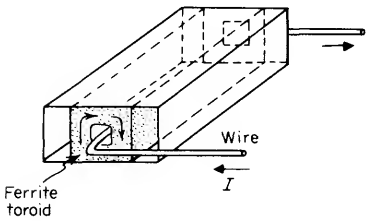


FIG 17-20 Waveguide latching digital phase shifter.

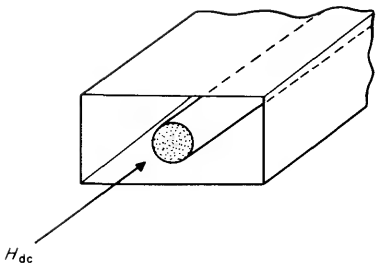


FIG 17-21 Reciprocal phase shifter.

17-4 Two-port Network Theory

This section presents a brief review of two-port network theory as an introduction to instruments and measurement techniques used to characterize microwave networks.

A linear network, or a nonlinear network operating with signals sufficiently small to cause the network to respond to a linear manner, can be described by parameters measured at the network terminals (ports) without regard to the contents of the network. Once the parameters of the network have been determined, its behavior again in any external environment can be predicted without regard to the specific content of the network.

Although a network may have any number of ports, network parameters can be explained most easily by considering a network with only two ports, an input port and an output port, like the network shown in Fig. 17-22. To define the performance of such a network, any of several parameter sets can be used, each of which has certain advantages.

Each parameter set contains four variables associated with the terminals of the two-port model. Two of these variables represent the exci-

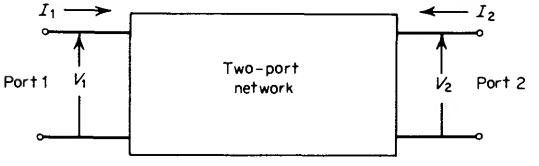


FIG 17-22 Two-port network.

tation of the network (independent variables), and the remaining two represent the response of the network to the excitation (dependent variables). These independent and dependent variables are related by four parameters describing the internal characteristics of the network. If the

network of Fig. 17-22 is driven by voltage sources V_1 and V_2, the terminal currents I_1 and I_2 will be determined by the following equations:

$$I_1 = y_{11}V_1 + y_{12}V_2 \tag{17-4-1}$$
$$I_2 = y_{21}V_1 + y_{22}V_2 \tag{17-4-2}$$

where V_1 and V_2 are the independent variables, I_1 and I_2 are the dependent variables, and y_{11}, y_{12}, y_{21}, and y_{22} are the network parameters describing the internal characteristics of the network.

In this case the network parameters are called *short-circuit admittance parameters* or *y parameters*. Four measurements are required to determine the four y parameters. Each measurement is made with one port of the network excited by a voltage source, while the other port is short circuited. For example, y_{21}, the forward transadmittance, is the ratio of the current at port 2 to the voltage at port 1 with port 2 short-circuited as shown in Eq. (17-4-3),

$$y_{21} = \frac{I_2}{V_1} \qquad V_2 = 0 \text{ (output short-circuited)} \tag{17-4-3}$$

If other independent and dependent variables had been chosen, the network would have been described, as before, by two linear equations similar to Eqs. (17-4-1) and (17-4-2), except the network parameters describing their relationships would have been different. The h, z, g, and $ABCD$ parameters are examples of other commonly used parameter sets. However, all parameter sets contain the same information about a network, and it is always possible to calculate any set in terms of any other set. The independent and dependent variables need not be the voltages and currents at the ports of the network as they are in the examples listed above. For example, the incident and reflected voltages or currents could equally well serve as terminal variables, and network parameters (s parameters) describing their interrelationships could also be defined.

The s parameters are being used in microwave design because they are easier to measure and apply at high frequencies than other parameters. They are conceptually simple, analytically convenient, and capable of providing insight into a measurement or design problem.

Scattering parameters are well suited for describing transistors and other active devices. Measuring most other parameters calls for the input and output of the device to be successively opened and short-circuited. This is difficult to do at rf frequencies where lead inductance and capacitance make short and open circuits difficult to obtain. At higher frequencies these measurements typically require tuning stubs, separately adjusted at each measurement frequency, to reflect short- or

open-circuit conditions to the device terminals. A tuning stub shunting the input or output may cause a transistor to oscillate, which makes the measurement difficult or invalid. Usually s parameters are measured with the device imbedded between a 50-Ω load and source, and there is less chance for oscillations to occur.

Traveling waves, unlike terminal voltages and currents, do not vary in magnitude at points along a lossless transmission line. This means that the scattering parameters of the network can be measured at a reference plane located some distance from the actual ports of the network.

17-5 Theory of s Parameters

Generalized scattering-parameter theory will now be discussed. The generalized theory applied to a two-port network will be covered in the next section.

Generalized scattering parameters have been defined by K. Kurokawa [24]. He defines power waves incident on, and reflected from, the kth port of a linear network as a_k and b_k respectively (refer to Fig. 17-23). These waves are related by the scattering matrix $[s]$, which characterizes the network:

$$[b_k] = [s][a_k] \tag{17-5-1}$$

where

$$a_k \equiv \frac{V_k + Z_k I_k}{2\sqrt{|\text{Re } Z_k|}} \tag{17-5-2}$$

$$b_k \equiv \frac{V_k - Z_k^* I_k}{2\sqrt{|\text{Re } Z_k|}} \tag{17-5-3}$$

and V_k = total voltage at kth port
I_k = total current at kth port
Z_k = arbitrary reference impedance for kth port (where * denotes the complex conjugate)

Let us discuss the waves defined in Eqs. (17-5-2) and (17-5-3). In the description of low-frequency networks, the voltage and current at the terminals are generally chosen as the independent variables. However, one may equally well choose any linear transformation of variables as long as the transformation is not singular, i.e., as long as the inverse transformation exists. The waves defined by Eqs. (17-5-2) and (17-5-3) are the results of just one of an infinite number of such linear transformations.

If V_k and I_k are given, a_k and b_k are calculated by Eqs. (17-5-2) and (17-5-3). On the other hand, if a_k and b_k are given, V_k and I_k are obtained

from the following inverse transformation:

$$V_k = \frac{P_k}{\sqrt{|\text{Re } Z_k|}} (Z_k^* a_k + Z_k b_k) \tag{17-5-4}$$

$$I_k = \frac{P_k}{\sqrt{|\text{Re } Z_k|}} (a_k - b_k) \tag{17-5-5}$$

where

$$P_k \equiv \begin{cases} 1 & \text{when Re } Z_k > 0 \\ -1 & \text{when Re } Z_k < 0 \end{cases}$$

Thus any result in terms of one set of variables can easily be converted to the other set of variables.

The power-flow concept of s parameters is most important. We shall now investigate this property. Figure 17-23 shows a general linear source with internal impedance Z_k, where the terms V_k, I_k, and Z_k are labeled, and the source voltage is V_o. The voltage at the source terminals is given by

$$V_k = V_o - Z_k I_k$$

If we insert this expression into Eq. (17-5-2) and take the square of the magnitude, we have

$$a_k^2 = \frac{|V_o|^2}{4|\text{Re } Z_k|} \tag{17-5-6}$$

which is equivalent to the available power from the source with a Z_k internal impedance. We make the convention that $|a_k|^2$ is positive if Re $Z_k > 0$ and negative if Re $Z_k < 0$.

Next, consider $|a_k|^2 - |b_k|^2$. Direct substitution of Eqs. (17-5-2) and (17-5-3) in this expression gives

$$\text{Re } (V_k I_k^*) = P_k(|a_k|^2 - |b_k|^2) \tag{17-5-7}$$

The left-hand side of Eq. (17-5-7) expresses the power actually transferred from the source to the load.

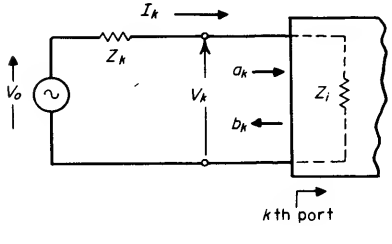


FIG 17-23 General source driving the Kth port of a network.

For a moment, consider the case where the real part of the internal impedance of the source is positive, that is, P_k is equal to 1. Then Eqs. (17-5-6) and (17-5-7) can be interpreted as follows: The source is sending the power $|a_k|^2$ toward the load, regardless of the load impedance. However, when the load is not matched, that is, Z_L does not equal Z_k^*, a part of the incident power is reflected back to the source. This reflected power is given by $|b_k|^2$, so that the net power absorbed in the load is equal to $|a_k|^2 - |b_k|^2$. Associated with these incident and reflected powers, there are waves a_k and b_k respectively.

For most measurements and calculations it is convenient to assume that the reference impedance Z_k is positive and real. When Z_k is real and positive, there is no difference in the expressions for the power waves of Eqs. (17-5-2) and (17-5-3) and the normalized traveling waves along a transmission line are defined in Eqs. (17-5-8) and (17-5-9),

$$a_k \equiv \frac{V_k + R_k I_k}{2\sqrt{R_k}} = \frac{V_{ik}}{\sqrt{R_k}} = \frac{\text{voltage incident on } k\text{th port}}{\sqrt{R_k}} \tag{17-5-8}$$

$$b_k \equiv \frac{V_k - R_k I_k}{2\sqrt{R_k}} = \frac{V_{rk}}{\sqrt{R_k}} = \frac{\text{voltage reflected from } k\text{th port}}{\sqrt{R_k}} \tag{17-5-9}$$

where we have used the following results from transmission-line theory:

$$V_k = V_{ik} + V_{rk}$$
$$I_k = I_{ik} - I_{rk}$$

and

$$R_k = \frac{V_{ik}}{I_{ik}} = \frac{V_{rk}}{I_{rk}}$$

If all a's and b's in Eq. (17-5-1) are referenced to the *same* positive-real

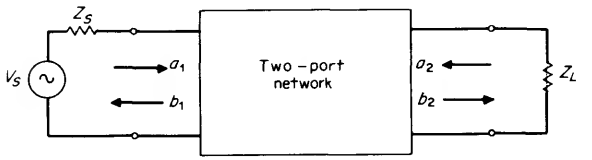


FIG 17-24 Two-port network with Z_s source and Z_L load.

impedance R_0, then we can divide both sides of the equation by $\sqrt{R_0}$ and obtain an expression relating incident and reflected voltages directly,

$$[V_{rk}] = [s][V_{ik}] \tag{17-5-10}$$

Two-port s-parameter Theory. We shall now apply the above theory to the two-port network shown in Fig. 17-24. For the two-port network,

Eq. (17-5-1) reduces to

$$b_1 = s_{11}a_1 + s_{12}a_2 \tag{17-5-11}$$
$$b_2 = s_{21}a_1 + s_{22}a_2 \tag{17-5-12}$$

where

$$a_1 = \frac{V_1 + Z_1 I_1}{\sqrt{|\text{Re } Z_1|}} \qquad a_2 = \frac{V_2 + Z_2 I_2}{\sqrt{|\text{Re } Z_2|}}$$
$$b_1 = \frac{V_1 - Z_1^* I_1}{\sqrt{|\text{Re } Z_1|}} \qquad b_2 = \frac{V_2 - Z_2^* I_2}{\sqrt{|\text{Re } Z_2|}}$$

and

$$s_{11} = \frac{b_1}{a_1} \qquad a_2 = 0$$
$$s_{22} = \frac{b_2}{a_2} \qquad a_1 = 0$$
$$s_{21} = \frac{b_2}{a_1} \qquad a_2 = 0$$
$$s_{12} = \frac{b_1}{a_2} \qquad a_1 = 0$$

However, if we reference both ports to R_0 (typically the characteristic impedance of a transmission line), Eq. (17-5-10) applies and Eqs. (17-5-4) and (17-5-12) simplify to the voltage-wave-scattering parameter equations

$$V_{r1} = s_{11}V_{i1} + s_{12}V_{i2} \tag{17-5-13}$$
$$V_{r2} = s_{21}V_{i1} + s_{22}V_{i2} \tag{17-5-14}$$

where

$$s_{11} = \frac{V_{r1}}{V_{i1}} \qquad V_{i2} = 0$$ input reflection coefficient with the output port terminated by a matched load ($Z_L = R_0$ sets $V_{i2} = 0$) (17-5-15)

$$s_{22} = \frac{V_{r2}}{V_{i2}} \qquad V_{i1} = 0$$ output reflection coefficient with the input terminated by a matched load ($Z_s = R_0$ and $V_s = 0$) (17-5-16)

$$s_{21} = \frac{V_{r2}}{V_{i1}} \qquad V_{i2} = 0$$ forward transmission gain with the output terminated in a matched load (17-5-17)

$$s_{12} = \frac{V_{r1}}{V_{i2}} \qquad V_{i1} = 0$$ reverse transmission gain with the input port terminated in a matched load. (17-5-18)

Equations (17-5-13) through (17-5-18) are very important and practical since the s parameters can be simply determined by measuring the incident and reflected voltages from the network.

Notice that

$$s_{11} = \frac{V_{r1}}{V_{i1}} = \frac{V_1 - R_0 I_1}{V_1 + R_0 I_1} = \frac{V_1/I_1 - R_0}{V_1/I_1 + R_0} = \frac{Z_i - R_0}{Z_i + R_0} \qquad V_{i2} = 0 \tag{17-5-19}$$

and

$$Z_i = R_0 \frac{(1 + s_{11})}{(1 - s_{11})} \tag{17-5-20}$$

where $Z_i = V_1/I_1$ is the input impedance at port 1.

This relationship between reflection coefficient and impedance is the basis of the Smith-chart transmission-line calculator. Consequently, the reflection coefficients s_{11} and s_{22} can be plotted on Smith charts and converted directly to impedance. This is just one example of how s parameters can be converted to another form. Table 17-2 is a listing of transformations between the common network parameter sets and s parameters for the two-port network.

From Fig. 17-23 and Eqs. (17-5-8) and (17-5-9), we see that the a and b coefficients (normalized to R_0) have physical significance. That is,

$$|a_1|^2 = \frac{|V_{i1}|^2}{R_0} = \begin{cases}\text{power incident on the input of the network, or} \\ \text{power available from a source of impedance } R_0\end{cases}$$

$$|a_2|^2 = \frac{|V_{i2}|^2}{R_0} = \begin{cases}\text{power incident on the output of the network, or} \\ \text{power reflected from the load}\end{cases}$$

$$|b_2|^2 = \frac{|V_{r1}|^2}{R_0} = \begin{cases}\text{power reflected from the input port of the network,} \\ \text{or power available from an } R_0 \text{ source minus the} \\ \text{power delivered to the input of the network}\end{cases}$$

$$|b_2|^2 = \frac{|V_{r2}|^2}{R_0} = \begin{cases}\text{power reflected or emanating from the output of} \\ \text{the network, power incident on the load, or power} \\ \text{that would be delivered to an } R_0 \text{ load.}\end{cases}$$

Hence, s parameters are related to power gain and mismatch loss, quantities which are of much interest, or

$$|s_{11}|^2 = \frac{\text{power reflected from the network input}}{\text{power incident on the network input}}$$

$$|s_{22}|^2 = \frac{\text{power reflected from the network output}}{\text{power incident on the network output}}$$

$$|s_{21}|^2 = \frac{\text{power delivered to an } R_0 \text{ load}}{\text{power available from } R_0 \text{ source}}$$

$$= \text{transducer power gain with } R_0 \text{ load and source}$$

$$|s_{12}|^2 = \text{reverse transducer power gain with } R_0 \text{ load and sources}$$

TABLE 17-2 Parameter Interrelationships†

$s_{11} = \dfrac{(z_{11} - 1)(z_{22} + 1) - z_{12}z_{21}}{(z_{11} + 1)(z_{22} + 1) - z_{12}z_{21}}$	$z_{11} = \dfrac{(1 + s_{11})(1 - s_{22}) + s_{12}s_{21}}{(1 - s_{11})(1 - s_{22}) - s_{12}s_{21}}$
$s_{12} = \dfrac{2z_{12}}{(z_{11} + 1)(z_{22} + 1) - z_{12}z_{21}}$	$z_{12} = \dfrac{2s_{12}}{(1 - s_{11})(1 - s_{22}) - s_{12}s_{21}}$
$s_{21} = \dfrac{2z_{21}}{(z_{11} + 1)(z_{22} + 1) - z_{12}z_{21}}$	$z_{21} = \dfrac{2s_{21}}{(1 - s_{11})(1 - s_{22}) - s_{12}s_{21}}$
$s_{22} = \dfrac{(z_{11} + 1)(z_{22} - 1) - z_{12}z_{21}}{(z_{11} + 1)(z_{22} + 1) - z_{12}z_{21}}$	$z_{22} = \dfrac{(1 + s_{22})(1 - s_{11}) + s_{12}s_{21}}{(1 - s_{11})(1 - s_{22}) - s_{12}s_{21}}$
$s_{11} = \dfrac{(1 - y_{11})(1 + y_{22}) + y_{12}y_{21}}{(1 + y_{11})(1 + y_{22}) - y_{12}y_{21}}$	$y_{11} = \dfrac{(1 + s_{22})(1 - s_{11}) + s_{12}s_{21}}{(1 + s_{11})(1 + s_{22}) - s_{12}s_{21}}$
$s_{12} = \dfrac{-2y_{12}}{(1 + y_{11})(1 + y_{22}) - y_{12}y_{21}}$	$y_{12} = \dfrac{-2s_{12}}{(1 + s_{11})(1 + s_{22}) - s_{12}s_{21}}$
$s_{21} = \dfrac{-2y_{21}}{(1 + y_{11})(1 + y_{22}) - y_{12}y_{21}}$	$y_{21} = \dfrac{-2s_{21}}{(1 + s_{11})(1 + s_{22}) - s_{12}s_{21}}$
$s_{22} = \dfrac{(1 + y_{11})(1 - y_{22}) + y_{21}y_{12}}{(1 + y_{11})(1 + y_{22}) - y_{12}y_{21}}$	$y_{22} = \dfrac{(1 + s_{11})(1 - s_{22}) + s_{12}s_{21}}{(1 + s_{22})(1 + s_{11}) - s_{12}s_{21}}$
$s_{11} = \dfrac{(h_{11} - 1)(h_{22} + 1) - h_{12}h_{21}}{(h_{11} + 1)(h_{22} + 1) - h_{12}h_{21}}$	$h_{11} = \dfrac{(1 + s_{11})(1 + s_{22}) - s_{12}s_{21}}{(1 - s_{11})(1 + s_{22}) + s_{12}s_{21}}$
$s_{12} = \dfrac{2h_{12}}{(h_{11} + 1)(h_{22} + 1) - h_{12}h_{21}}$	$h_{12} = \dfrac{2s_{12}}{(1 - s_{11})(1 + s_{22}) + s_{12}s_{21}}$
$s_{21} = \dfrac{-2h_{21}}{(h_{11} + 1)(h_{22} + 1) - h_{12}h_{21}}$	$h_{21} = \dfrac{-2s_{21}}{(1 - s_{11})(1 + s_{22}) + s_{12}s_{21}}$
$s_{22} = \dfrac{(1 + h_{11})(1 - h_{22}) + h_{12}h_{21}}{(h_{11} + 1)(h_{22} + 1) - h_{12}h_{21}}$	$h_{22} = \dfrac{(1 - s_{22})(1 - s_{11}) - s_{12}s_{21}}{(1 - s_{11})(1 + s_{22}) + s_{12}s_{21}}$

† The h, y, and z parameters listed are all normalized to Z_0. If h', y', and z' are the actual parameters, conversion is obtained as follows:

$$z'_{11} = z_{11}Z_0 \qquad y'_{11} = \frac{y_{11}}{Z_0} \qquad h'_{11} = h_{11}Z_0$$

$$z'_{12} = z_{12}Z_0 \qquad y'_{12} = \frac{y_{12}}{Z_0} \qquad h'_{12} = h_{12}$$

$$z'_{21} = z_{21}Z_0 \qquad y'_{21} = \frac{y_{21}}{Z_0} \qquad h'_{21} = h_{21}$$

$$z'_{22} = z_{22}Z_0 \qquad y'_{22} = \frac{y_{22}}{Z_0} \qquad h'_{22} = \frac{h_{22}}{Z_0}$$

17-6 Network Calculations by Using Scattering Parameters

Scattering parameters are convenient in many network calculations. This is especially true for power and power-gain calculations. For example, the transfer parameters s_{12} and s_{21} are a measure of the complex insertion gain, and the driving-point parameters s_{11} and s_{22} are a measure

of the input and output mismatch loss. The s parameters form a natural set for use with signal-flow graphs. Of course, it is not necessary to use signal-flow graphs in order to use s parameters.

In s-parameter signal-flow graphs, each port is represented by two nodes. Node a_k represents the wave coming into the network at port k, and b_k represents the wave leaving the network at port k. The complex scattering coefficients are represented by multipliers on directed branches connecting the nodes within the network. The transfer function between any two nodes in the network can be determined by using topological flow-graph reduction techniques or by applying Mason's "nontouching-loop rule."

The complete flow graph of the network in Fig. 17-24, including source and load impedances, is shown in Fig. 17-25. The load and source are described by their reflection coefficients Γ_L and Γ_S respectively, referenced to the real characteristic impedance R_0. Notice that b_0 represents the normalized voltage delivered from a source with a Z_s internal impedance to an R_0 termination and that $|b_0|^2$ equals the power dissipated in an R_0 load.

We shall now explain the topological flow-graph reduction technique. The solution for the flow graph can be arrived at by a series of topological manipulations which reduce a flow graph to simpler and simpler forms until the answer is obvious. The manipulations are based on the following four rules:

Rule 1. Two branches, whose common node has only one incoming and one outgoing branch (branches in series), may be combined to form a single branch whose coefficient is the product of the coefficients of the

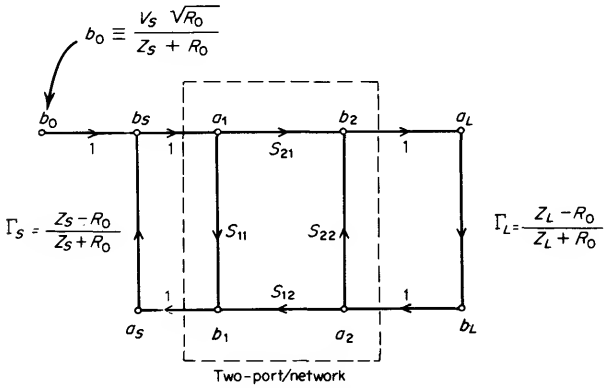


FIG 17-25 An s-parameter flow graph of two-port network with Z_s source and Z_L load.

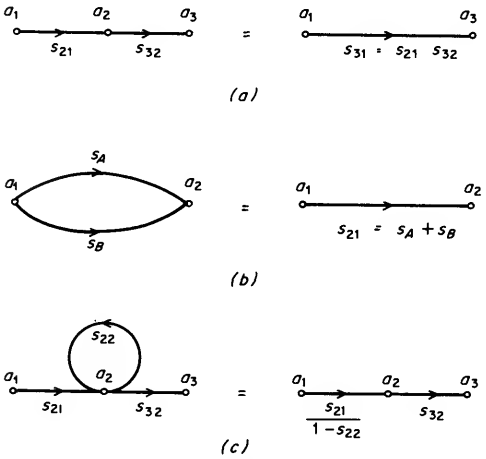


FIG 17-26 Reduction rules: (*a*) branches in series, (*b*) branches in parallel, (*c*) reduction of a self-loop.

original branches. Thus the common node is eliminated (see Fig. 17-26*a*).

Rule 2. Two branches pointing from a common node to another common node (branches in parallel) may be combined into a single branch whose coefficient is the sum of the coefficients of the original branches (see Fig. 17-26*b*).

Rule 3. When a node k possesses a self-loop (a branch which begins and ends at k), the self-loop may be eliminated by dividing the coefficient of every other branch entering node k by 1 − (coefficient of self-loop). Figure 17-26*c* shows the elimination of a self-loop by this rule.

Rule 4. A node may be duplicated, that is, split into two nodes which may be subsequently treated as two separate nodes, so long as the original signal flows are not altered. Any self-loop attached to the original node must also be attached to *each* of the nodes resulting from duplication. Figure 17-27 illustrates various node-splitting techniques.

Mason's nontouching loop rule will now be discussed. The nontouching-loop rule provides a method for writing the solution of any flow graph by inspection. The solution T (the ratio of the output variable to the input variable) is

$$T = \frac{\sum_k T_k \Delta_k}{\Delta} \tag{17-6-1}$$

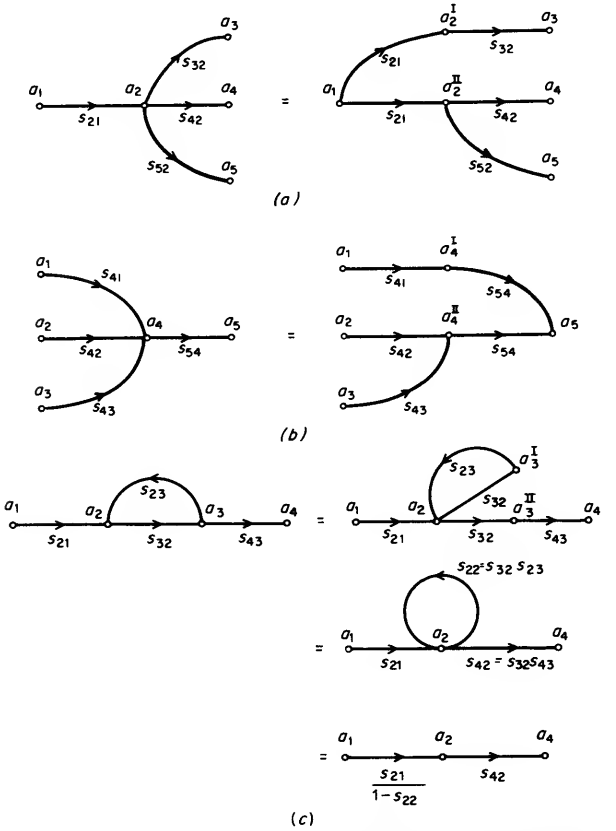


FIG 17-27 Node duplication rule: (*a*) node with single input branch, (*b*) node with single output branch, (*c*) node duplication with a feedback loop to obtain a self-loop.

where T_k = path gain of kth forward path

Δ = 1 − (sum of all first-order loop gains) + (sum of all second-order loop gains) − (sum of all third-order loop gains)

Δ_k = 1 − (sum of all first-order loop gains not touching the kth forward path) + (sum of all second-order loop gains not touching the kth forward path) − · · ·

A path is a continuous succession of branches. A forward path is a path connecting the input node to the output node, where no node is encountered more than once. Path gain is the product of all the branch multipliers along the path. A first-order loop is a path that originates and

terminates on the same node, no node being encountered more than once. A second-order loop is the product of two nontouching first-order loops (that is, two first-order loops that have no branches or nodes in common). A third-order loop is the product of three nontouching first-order loops, etc. Loop gain is the product of the branch multipliers around the loop.

As examples of using scattering-parameter flow graphs and the nontouching-loop rule, we shall calculate the network functions that follow.

Input Reflection Coefficient with Arbitrary Output Termination. Let us first solve this example by flow-graph reduction techniques. The flow graph is shown in Fig. 17-28*a*; the various steps used to reduce the flow graph are illustrated in Fig. 17-28*b* through *e*. Figure 17-28*b* is obtained by applying Rule 4 and duplicating node a_2. Rule 3 is used to reduce the self-loop at node b_2 to the form shown in Fig. 17-28*c*. Then by applying Rules 1 and 2 we obtain the final results shown in Fig. 17-28*a*.

The same result can be achieved by applying the nontouching-loop

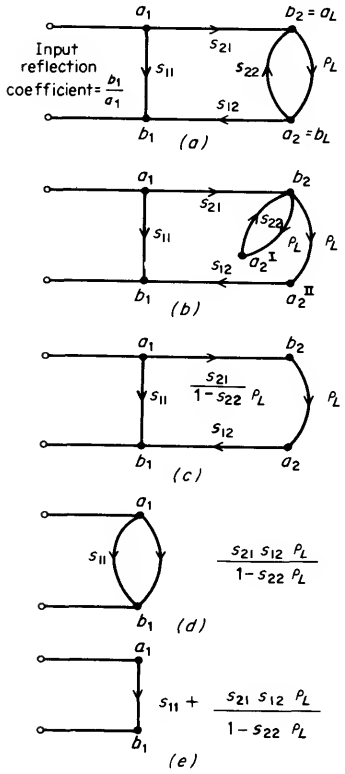


FIG 17-28 Flow-graph reduction to obtain the input reflection coefficient with arbitrary ρ_L.

rule. First we locate the forward paths from node a_1 to node b_1, using Fig. 17-28*a*:

Forward path 1 = s_{11}
Forward path 2 = $s_{21}\Gamma_L s_{12}$

Next we find all the first-order loops (for this example there is only one):

Sum of all first-order loop gains = $s_{22}\Gamma_L$

There are no higher-order loops in this example. Thus, for this example,

$$\Delta = 1 - s_{22}\Gamma_L$$

The first forward path s_{11} has one loop which does not touch it, and therefore, Δ_k (with $k = 1$) is

$$\Delta_1 = 1 - s_{22}\Gamma_L$$

with $k = 2$, $\Delta_k = \Delta_2 = 1$.

We can now write the answer:

$$T = \frac{b_1}{a_1} = \frac{s_{11}(1 - s_{22}\Gamma_L) + s_{21}s_{12}\Gamma_L(1)}{1 - s_{22}\Gamma_L}$$

$$\frac{b_1}{a_1} = s_{11} + \frac{s_{21}s_{12}\Gamma_L}{1 - s_{22}\Gamma_L} \tag{17-6-2}$$

Voltage Gain with Arbitrary Load Impedance. In this example,

$$K_V = \frac{V_2}{V_1}$$
$$V_1 = (a_1 + b_1)\sqrt{R_0} = V_{i1} + V_{r1}$$
$$V_2 = (a_2 + b_2)\sqrt{R_0} = V_{i2} + V_{r2}$$
$$a_2 = \Gamma_L b_2$$
$$b_1 = s'_{11}a_1$$

and so

$$K_V = \frac{b_2}{a_1}\frac{1 + \Gamma_L}{1 + s'_{11}}$$

Using the nontouching-loop rule,

$$K_V = \frac{s_{21}(1 + \Gamma_L)}{(1 - s_{22}\Gamma_L)(1 + s'_{11})} \tag{17-6-3}$$

Power Available from the Source. The flow graph for the source with an arbitrary load Γ_L is shown in Fig. 17-29. The power absorbed by the load is $|a_2|^2 - |b_2|^2$. The solution of the flow graph for a_2 and b_2 in terms

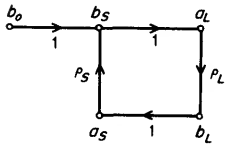


FIG 17-29 Flow graph used to determine the available power from a source with internal impedance Z_s.

of b_0 is as follows:

$$a_L = b_0 + \Gamma_L\Gamma_S a_L$$

$$a_L = \frac{b_0}{1 - \Gamma_L\Gamma_S}$$

$$b_L = \Gamma_L a_L = \frac{\Gamma_L b_0}{1 - \Gamma_L\Gamma_S}$$

$$|a_L|^2 - |b_L|^2 = \frac{|b_0|^2(1 - |\Gamma_L|^2)}{|1 - \Gamma_L\Gamma_S|^2}$$

The maximum power absorbed by the load (P_{avs}) occurs when $\Gamma_L = \Gamma_S^*$; therefore

$$(|a_2|^2 - |b_2|^2)\ \max \equiv P_{avs} = \frac{|b_0|^2}{1 - |\Gamma_S|^2} \tag{17-6-4}$$

Transducer Power Gain with Arbitrary Load and Source Impedances

$$G_T \equiv \frac{\text{power delivered to the load}}{\text{power available from the source}} = \frac{P_L}{P_{avs}}$$

$$P_L = P\ (\text{incident on load}) - P\ (\text{reflected from load})$$
$$= |b_2|^2(1 - |\Gamma_L|^2)$$

$$P_{avs} = \frac{|b_0|^2}{(1 - |\Gamma_S|^2)}$$

$$G_T = \left|\frac{b_2}{b_0}\right|^2 (1 - |\Gamma_S|^2)(1 - |\Gamma_L|^2)$$

From the nontouching-loop rule,

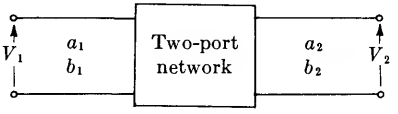
$$\frac{b_2}{b_0} = \frac{s_{21}}{1 - s_{11}\Gamma_S - s_{22}\Gamma_L - s_{21}s_{12}\Gamma_L\Gamma_S + s_{11}\Gamma_S s_{22}\Gamma_L}$$

$$= \frac{s_{21}}{(1 - s_{11}\Gamma_S)(1 - s_{22}\Gamma_L) - s_{21}s_{12}\Gamma_L\Gamma_S}$$

$$G_T = \frac{|s_{21}|^2(1 - |\Gamma_S|^2)(1 - |\Gamma_L|^2)}{|(1 - s_{11}\Gamma_S)(1 - s_{22}\Gamma_L) - s_{21}s_{12}\Gamma_L\Gamma_S|^2} \tag{17-6-5}$$

Table 17-3 lists formulas for calculating many often-used network functions (power gains, driving-point characteristics, etc.) in terms of scattering parameters. Not only are scattering-parameter flow graphs helpful in solving various network calculations, but they are also useful in determining measurement errors and uncertainties.

TABLE 17-3 Network Calculations by Using s Parameters



$$b_1 = s_{11}a_1 + s_{12}a_2$$
$$b_2 = s_{21}a_1 + s_{22}a_2$$

Input reflection coefficient with arbitrary Z_L

$$s'_{11} = s_{11} + \frac{s_{21}s_{12}\Gamma_L}{1 - s_{22}\Gamma_L}$$

Output reflection coefficient with arbitrary Z_s

$$s'_{22} = s_{22} + \frac{s_{12}s_{21}\Gamma_s}{1 - s_{11}\Gamma_s}$$

Voltage gain with arbitrary Z_L

$$K_v = \frac{V_2}{V_1} = \frac{s_{21}(1 + \Gamma_L)}{(1 - s_{22}\Gamma_L)(1 + s'_{11})}$$

$$\text{Power gain} \equiv \frac{\text{power delivered to load}}{\text{power input to network}}$$

$$G = \frac{|s_{21}|^2(1 - |\Gamma_L|^2)}{(1 - |s_{11}|^2) + |\Gamma_L|^2(|s_{22}|^2 - |D|^2) - 2\,\mathrm{Re}\,(\Gamma_L N)}$$

$$\text{Available power gain} \equiv \frac{\text{power available from network}}{\text{power available from source}}$$

$$G_A = \frac{|s_{21}|^2(1 - |\Gamma_s|^2)}{(1 - |s_{22}|^2) + |\Gamma_s|^2(|s_{11}|^2 - |D|^2) - 2\,\mathrm{Re}\,(\Gamma_s M)}$$

$$\text{Transducer power gain} \equiv \frac{\text{power delivered to load}}{\text{power available from source}}$$

$$G_T = \frac{|s_{21}|^2(1 - |\Gamma_s|^2)(1 - |\Gamma_L|^2)}{|(1 - s_{11}\Gamma_s)(1 - s_{22}\Gamma_L) - s_{12}s_{21}\Gamma_L\Gamma_s|^2}$$

where $D \equiv s_{11}s_{22} - s_{12}s_{21}$
$M \equiv s_{11} - Ds_{22}$
$N \equiv s_{22} - Ds_{11}$

17-7 Measurement of *s* Parameters

In this section, assume that the characteristic impedance of the transmission lines at *all* ports equals R_0 so that Eqs. (17-5-13) through (17-5-18) apply. Therefore measurement of the incident, reflected, and transmitted *voltages*, with the ports properly terminated, will be sufficient to determine the network *s* parameters.

Since $|s_{11}|$ and $|s_{22}|$ are simply the magnitude of the reflection coefficients of each port of the network with the other port terminated in an R_0 load, $|s_{11}|$ and $|s_{22}|$ can easily be determined by using the reflectometer method described earlier. If we want the complex reflection coefficients s_{11} and s_{22}, the measurement system used for determining the driving-point impedance of a network or a slotted line can be used. Because $|s_{21}|$ is the insertion loss (or gain) of the two-port network, it can be measured by the techniques described in the section on measurement of attenuation. The above techniques have been discussed in detail earlier and will not be reviewed here. It is important to realize that the above classical methods are valid techniques for measuring *s* parameters and will suffice for many of our measurement needs. However, there are times when we want complete device characterization. Both the magnitude and the phase of all four *s* parameters (in the two-port network) will sometimes be required. This section will describe *one* measurement system used for complete network characterization.

The Network-analyzer Concept. The *network-analyzer* system shown in Fig. 17-30 contains three main functions: signal source, test set, and receiver. The signal source can be either continuous-wave signal generator or a sweeper. The sweeper can be used for broadband network characterization. The amplitude of the swept rf output may or may not

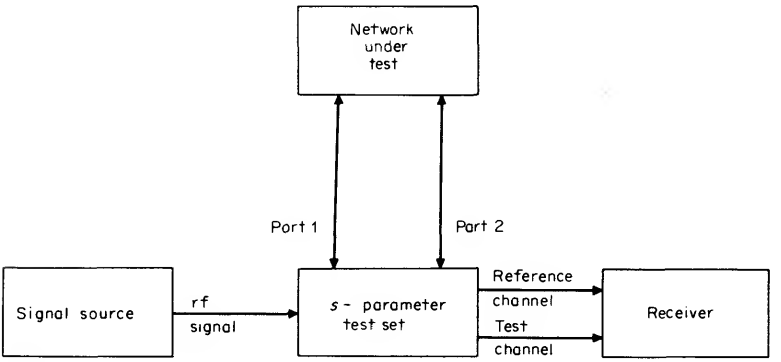


FIG 17-30 Network-analyzer system.

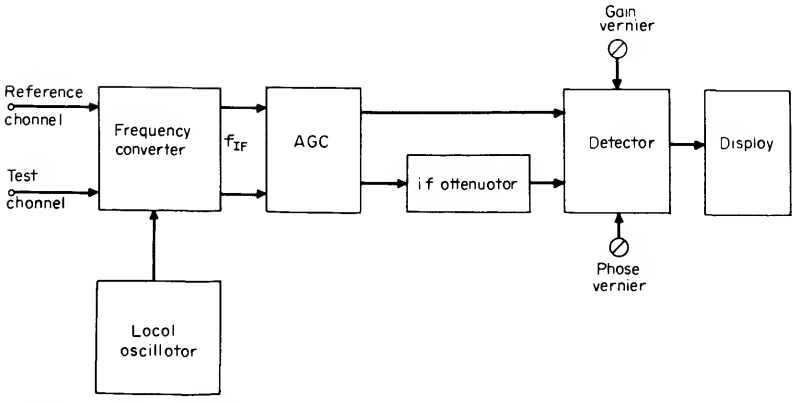


FIG 17-31 Receiver block diagram.

need to be leveled, depending on whether or not the receiver has an automatic gain control.

The *s*-parameter test set routes rf power to the network under test and separates the incident, reflected, and transmitted signals needed to characterize the network *s* parameters. The incident signal is routed to the reference channel of the receiver, and either the reflected or transmitted signals can be connected to the test channel of the receiver. An adjustable "line stretcher" is provided in the test set to equalize the apparent electrical lengths of the reference and test channels.

The receiver is a two-channel vector-ratio detector; it measures the amplitude ratio and phase difference between the reference and test channels. A block diagram of the receiver is shown in Fig. 17-31. Observe that the broadband input signals are down-converted to a constant intermediate frequency f_{IF}. This down-conversion is a linear process and does not alter the phase and amplitude information of the original rf signals. Thus, gain and phase information are preserved, and all signal processing and measurements can take place at a constant frequency. The local oscillator can be synchronized to keep the receiver automatically tuned (phase-locked) to the rf frequency from the signal source. The automatic gain control normalizes the test channel amplitude to the reference channel amplitude. Any amplitude changes that are common to both reference and test channels are thus eliminated in effect, and we need measure only the amplitude of the test channel to determine the amplitude ratio.

The detection process can be performed in either of two ways: (1) Measure the amplitude of the test channel and the phase difference between reference and test channels, or (2) measure the real and imaginary

components of the test channel normalized to the reference channel phase. The first method provides the amplitude ratio and phase difference data convenient for displaying transmission phase and gain (or attenuation). Usually the log of the amplitude ratio is displayed. The second method provides data in polar form, which is convenient for displaying reflection coefficients (Smith chart). The intermediate-frequency attenuator is used as an amplitude offset control when the log of the amplitude ratio is displayed. When the polar display is used, the intermediate-frequency attenuator expands or contracts the magnitude of the polar vector. Phase and gain verniers are provided for calibrating the displays. An external scope, *xy* recorder, or voltmeter can also be used to display the data from the detectors.

17-8 Measurement Techniques

When making microwave measurements, one should not blindly rush in and take the measurement and hope that results will be accurate. A four-step procedure that should be followed to improve measurement technique follows:

1. Error analysis: Analyze the measurement system errors.
2. Error reduction: Reduce the major system errors.
3. Initialization: Calibrate the system for the measurement.
4. Measurement: Insert and measure the unknown network.

Let us discuss these four steps in detail.

Error analysis is important in helping us to recognize the limitations of our measurement system. We know that low directivity in a coupler can cause measurement errors; also tracking between coupler arms, mismatch errors, frequency tracking of the frequency converter, calibration errors in the attenuator, noise, etc., can cause measurement ambiguities. The importance of these error terms depends on the level of accuracy we demand in our particular measurement situation. If we want to measure to an accuracy within 1 dB, we can just about make the measurement any way we want and be assured of good results; but if we want 0.1-dB accuracy, we need to examine the magnitude of the various error terms and take steps to reduce some of them. For 0.01-dB accuracy we not only need to know the magnitude of the error terms, but may need the phase of these terms as well. It will become necessary to use error-reduction techniques to remove many of the error terms. When making measurements with high accuracy, we may not be able to remove the errors physically; but if we know and remember what they are, we can take account of them mathematically in our final measured data. Error-reduction techniques will be covered later when we deal with specific measurement examples.

We shall now develop the *error models* needed to characterize the error terms in our measurement system. For *any* two-port *s*-parameter measurement system, Table 17-4 lists the possible error parameters and their physical interpretation. Figure 17-32*a* through *d* shows the four basic error models with the use of these parameters.

The general error-analysis procedure is to write the flow graph of the entire system and then solve for the appropriate transfer function by using the flow-graph reduction technique or the nontouching-loop rule. This can become difficult for a measurement system like that shown in Fig. 17-30. However, these error models can be simplified in specific measurement situations after reducing the models to the basic forms of Fig. 17-32. The error models and error parameters will be described in detail when we discuss specific measurement examples.

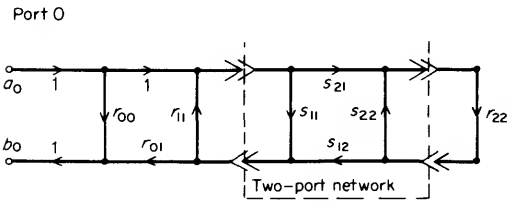
Assume that the error terms and ambiguities are now accounted for and minimized. We are ready to initialize the measurement system. During the initialization we may equalize the electrical length in the reference and test channels and adjust the phase and amplitude vernier controls to scale our display, etc. We are now ready to complete the measurement, but a final word of caution. Be sure all connectors are tight and that no undue strain is applied to the mechanical interconnections. This is extremely important at high microwave frequencies; faulty connections can cause hours of grief and inaccurate results.

Reflection Measurements. The system shown in Fig. 17-33*a* can be used for one-port reflection measurements. The basic block diagram of the system described in Fig. 17-30 is used, where the test set is a simple reflectometer. The line stretcher was added to the reflectometer to balance the electrical length in the reference and test channels. The error model, the equation describing the reflection measured by the

Table 17-4 Common Error Sources for the Error Models in Fig. 17-32

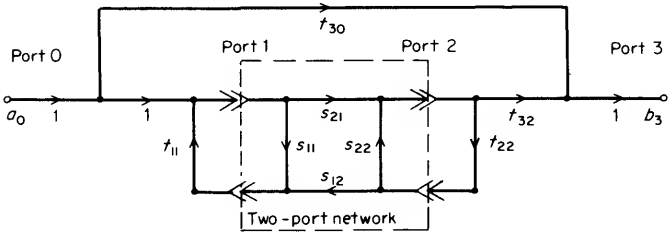
Error parameters†		Physical counterpart
r_{00}	r_{33}	Reflection crosstalk (directivity)
t_{30}	t_{03}	Transmission crosstalk
r_{01}	r_{32}	Reflection tracking errors
t_{32}	t_{01}	Transmission tracking errors
$r_{11}r_{22}$	$r'_{11}r'_{22}$	Reflection port mismatch
$t_{11}t_{22}$	$t'_{11}t'_{22}$	Transmission port mismatch

† To the above errors we should add the noise of the system, attenuator errors, detector errors, and display errors.



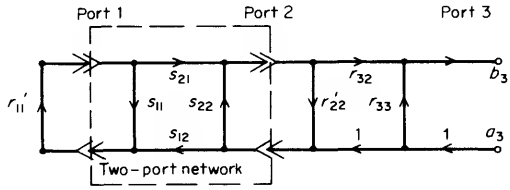
$$\text{Meosured } s_{11} \equiv \frac{b_0}{a_0} = r_{00} + \frac{s_{11}\, r_{01}(1 - s_{22} r_{22}) + s_{21}\, s_{12}\, r_{22}\, r_{01}}{1 - s_{11} r_{11} - s_{22} r_{22} - s_{21} s_{12} r_{11} r_{22} + s_{11} r_{11} s_{22} r_{22}}$$

(a)



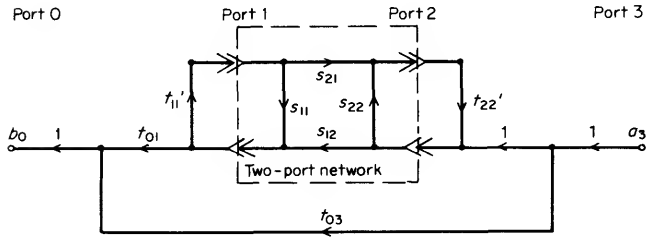
$$\text{Meosured } s_{21} \equiv \frac{b_3}{a_0} = t_{30} + \frac{s_{21}\, t_{32}}{1 - s_{11} t_{11} - s_{22} t_{22} - s_{21} s_{12} t_{11} t_{22} + s_{11} t_{11} s_{22} t_{22}}$$

(b)



$$\text{Meosured } s_{22} \equiv \frac{b_3}{a_3} = r_{33} + \frac{s_{22}\, r_{32}(1 - s_{11} r_{11}') + s_{12}\, s_{21}\, r_{11}'\, r_{32}}{1 - s_{22} r_{22}' - s_{11} r_{11}' - s_{12} s_{21} r_{22}' r_{11}' + s_{22} r_{22}' s_{11} r_{11}'}$$

(c)



$$\text{Meosured } s_{12} \equiv \frac{b_0}{a_3} = t_{03} + \frac{s_{12}\, t_{01}}{1 - s_{22} t_{22}' - s_{11} t_{11}' - s_{12} s_{21} t_{22}' t_{11}' + s_{22} t_{22}' s_{11} t_{11}'}$$

(d)

FIG 17-32 Measurement-system error models: (*a*) for measuring s_{11}, (*b*) for measuring s_{21}, (*c*) for measuring s_{22}, and (*d*) for measuring s_{12}. Ports 1 and 2 are the actual ports of the network analyzer; port 0 and port 3 are fictitious ports used only for notational purposes.

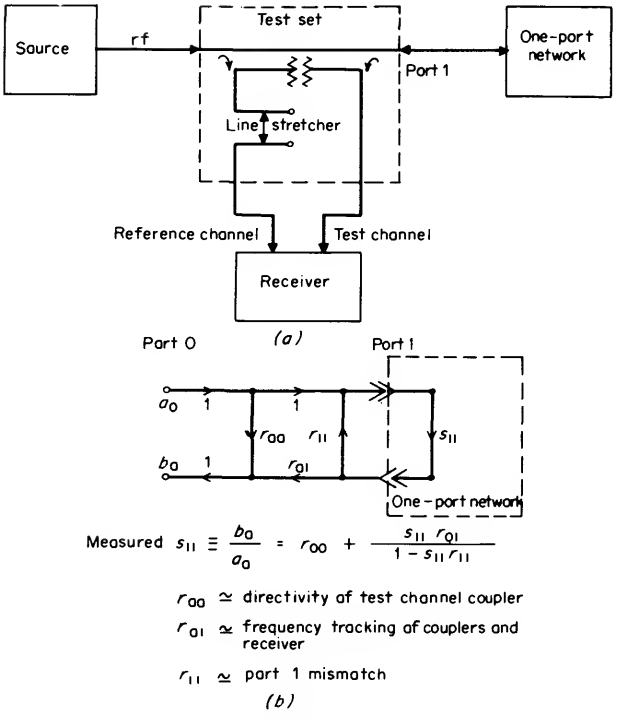


FIG 17-33 One-port measuring system: (*a*) block diagram, and (*b*) error model.

system, and a description of the various error sources are given in Fig. 17-33*b*. This error model is simplified from the model in Fig. 17-32*a* since $s_{12} = s_{21} = 0$ for the one-port network.

We shall now describe the various error terms and the sources from which they arise. If we place a perfect load on port 1 ($s_{11} = 0$), we still measure an apparent reflection. This may be caused by a signal leaking either from the reference channel to the test channel in the receiver, or via the directivity path through the test channel coupler. The leakage term will typically be down 80 dB or more below a full reflection, but the directivity will typically be no better than 40 dB below a full reflection. Hence, the directivity of the test channel coupler is the greatest contributor to the crosstalk term r_{00}. If the amplitude and phase response of the coupler arms do not track or the two channels of the receiver do not track, we measure an apparent ripple as we vary frequency, which cannot be attributed to the network being tested. This ripple term is known as

FIG 17-34 Coupler directivity calibration (r_{00}) obtained by phasing a sliding load through one wavelength at a single frequency. The directivity vector magnitude is the distance from the center of the display to the center of the small circle obtained when the load is phased through 360°. Fullscale $s_{11} = 0.05$.

the *frequency tracking error* r_{01}. Finally, we know that an imperfect port match r_{11} can cause multiple mismatch errors.

Since the directivity r_{00} is typically a small number, it can be neglected many times when measuring large values of s_{11}. However, when s_{11} is small, r_{11} becomes significant and *must* be reduced to obtain accurate measurements. The tracking error r_{01} causes a percentage error in the measured value of s_{11} and is sometimes neglected when s_{11} is small. Also the mismatch error r_{11} is not as important when s_{11} is small. The mismatch and tracking terms, r_{11} and r_{01}, affect the measurement of s_{11} most severely when s_{11} is large.

In the continuous-wave mode of operation the errors can be calibrated. First the source mismatch term r_{11} can be reduced by using a slide screw tuner, etc. Then a sliding load can be used to isolate the directivity. Attach the sliding load to the output port and slide the load back and forth (phasing the load). The resultant output on a polar display is shown in Fig. 17-34. The reflection coefficient of the sliding load is not zero, but when the load vector is rotated through 360°, its center can be located, which isolates the directivity vector r_{00}. To determine the tracking error r_{01}, place a short circuit on the output port. The measured reflection should be 1 at an angle of 180°. The phase and amplitude vernier controls can be adjusted to obtain the correct output on the display.

For swept measurements, we first need to equalize the electrical length

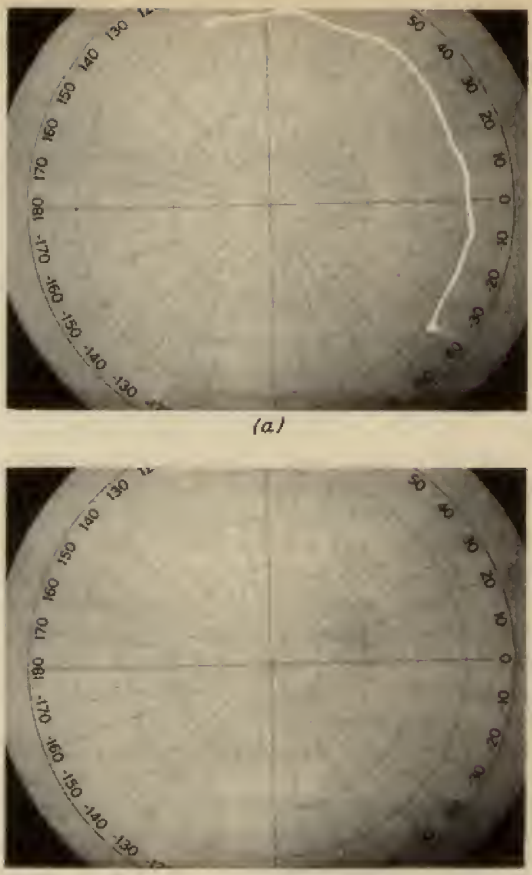


FIG 17-35 Equalizing electrical length in reflection measurement system: (*a*) swept measurement of a short before electrical length equalizing, and (*b*) result of equalization and adjusting phase and amplitude verniers.

in the reference and test channels. The length can be equalized by placing a short circuit on the output of the measurement system. Using a polar display, we obtain the results of Fig. 17-35. The line stretcher is adjusted until the line becomes a small cluster, and then the phase vernier control is adjusted until the display reads an angle of 180°. For continuous-wave measurements, the line stretcher does not need to be adjusted. For both continuous-wave and swept measurements, the line stretcher can be used to extend or retract the reference plane if we

want a different measurement plane from that established by the short.

In making swept measurements, the frequency tracking error can be approximately isolated by connecting a short to port 1 and drawing grid lines on the display device for each different setting of the intermediate-frequency attenuator as described in Fig. 17-36. If the mismatch term r_{11} causes significant measurement error, it may be reduced by inserting a pad between port 1 and the network under test.

In making reflection measurements on a two-port network, the same one-port measurement system can be used if we add a good termination to the output of the network, as shown in Fig. 17-37*a*. In this case, one additional error term must be considered; this is the mismatch of the termination placed on the output port of the network under test, r_{22}. Figure 17-37*b* shows the error model for this case. Notice that if s_{21} or s_{12} is small, we may be able to neglect this mismatch term; but if s_{21} and s_{12} are close to unity, we may need to tune out this mismatch, or at least reduce it by using a high-quality termination. Except for the above complication, the two-port reflection measurement proceeds as in the one-port system.

The above error discussions are based on a linear error model, but in the real world there are many other types of errors that can occur. For example, nonlinear errors in the receiver front end can cause gain compression errors or a phase that changed with amplitude changes. Connector repeatability is important. Many times the crosstalk changes with signal level changes or as the phase changes. Attenuators are not perfect and do cause phase and amplitude errors when they are used. Noise is always present, and so is Hiesenberg's uncertainty principle, and

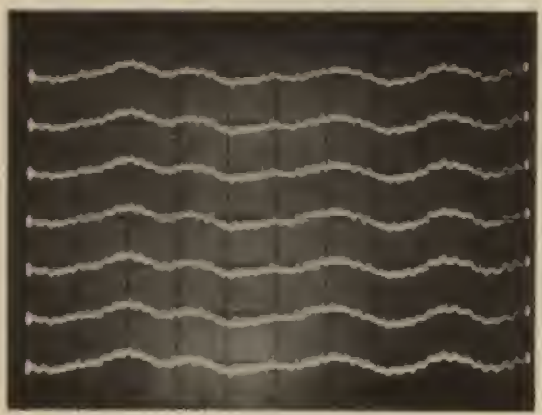
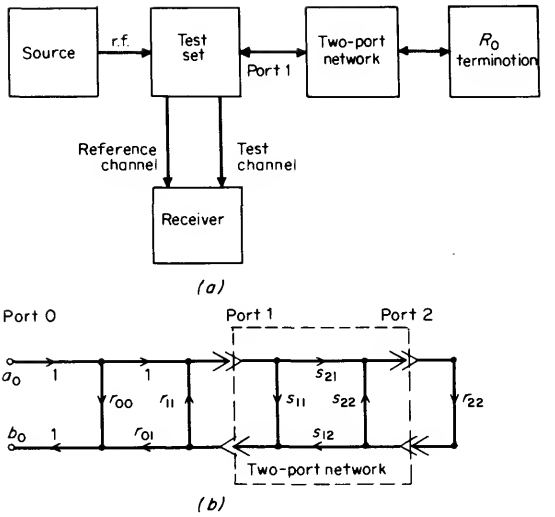


FIG 17-36 Tracking-error calibration grid lines for reflection measurements.



$$\text{Measured } s_{11} = \frac{b_0}{a_0} = r_{00} + \frac{s_{11}r_{01}(1 - s_{22}r_{22}) + s_{21}s_{12}r_{22}r_{01}}{(1 - s_{11}r_{11})(1 - s_{22}r_{22}) - s_{21}s_{12}r_{11}r_{22}}$$

$r_{00} \simeq$ directivity of test channel coupler
$r_{01} \simeq$ frequency tracking of couplers and receiver
$r_{11} \simeq$ port 1 mismatch
$r_{22} =$ termination mismatch

FIG 17-37 Two-port reflection measurement system: (*a*) block diagram, and (*b*) error model.

so forth. Most of these errors are difficult to calibrate and are often overlooked by the measurement system user.

Many times the input impedance of an active device is negative. This means the magnitude of the reflection coefficient is greater than unity. The standard Smith chart does not provide for $|s_{11}| > 1$, but the compressed Smith chart in Fig. 17-38 can be used whenever $|s_{11}| < 3.16$ and provides a direct readout of negative impedance.

Transmission Measurements. A typical system used for transmission measurements, along with the error model and mathematical interpretation, is shown in Fig. 17-39. This system compares the transmission through the reference channel with the test channel. The network tested is thus normalized to the reference channel path. If the amplitude and electrical length of the reference channel varies with frequency or the two arms of the power splitter do not track or the receiver channels do not track, we measure an apparent ripple t_{32} in our transmission measurement. The port mismatch terms t_{11} and t_{22} also cause measurement errors. If

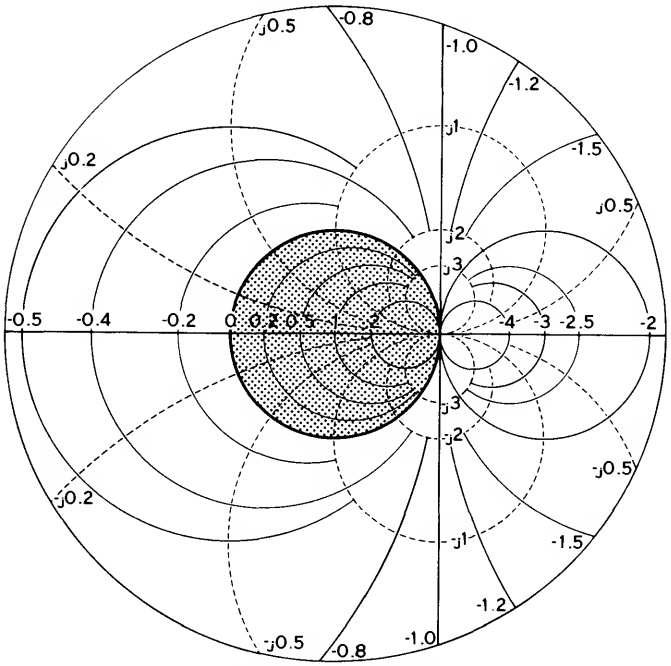


FIG 17-38 Compressed Smith chart used for direct readout of negative impedances. Full-scale $|s_{11}| = 3.16$.

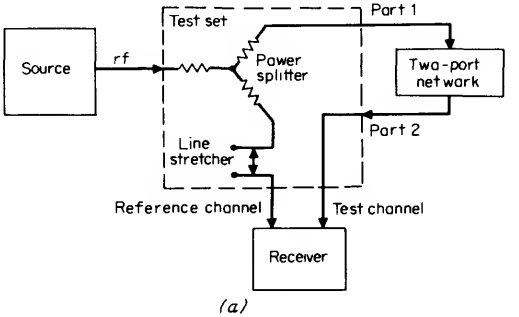
we do not connect any network, we still measure some transmission of the crosstalk in the system t_{30}.

Let us now discuss the importance of these various error terms. The crosstalk term t_{30} is usually a very small number. Typically it is down 80 dB or more and only causes problems in the measurement of high levels of attenuation. Many times the crosstalk is about the same level as the noise in the system. The tracking error t_{32} causes a percentage error in the transmission measurement. The mismatch terms t_{11} and t_{22} cause measurement error even when s_{11} and s_{22} of the network are zero, because t_{11} and t_{22} are coupled by the s_{21} and s_{12} paths of the network under test. The multiple mismatch terms $t_{11}s_{11}$ and $t_{22}s_{22}$ reduce as s_{11} and s_{22} approach zero and are many times neglected in measuring a network with low input and output reflection coefficients. It is difficult to measure the attenuation of a bilateral network with high input and output reflections.

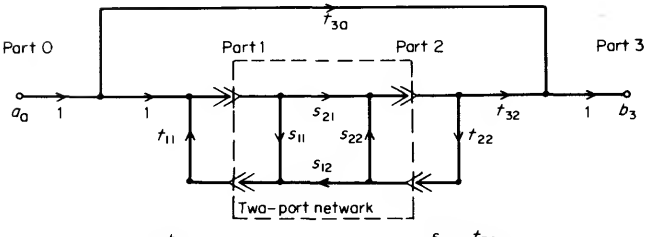
In the continuous-wave mode of operation the errors can be calibrated in much the same way as described in the reflection measurement section. First we tune out both mismatches. Then we disconnect ports 1 and 2

so that $s_{12} = s_{21} = 0$. This isolates the crosstalk term t_{30}. However, t_{30} is sometimes hard to recover from the noise. Next we connect ports 1 and 2 and adjust the gain and pulse vernier controls until the magnitude is unity and the phase is 0°. This calibrates the tracking term t_{32}.

For swept-transmission measurements the electrical length in the test and reference channels must first be equalized. This can be accomplished by the same technique used in the reflection measurement. For the transmission measurement we connect the two ports and adjust the line stretcher and phase vernier until the small cluster is centered at 0° on the polar display. A phase-versus-frequency display could equally well be used for adjusting the electrical length. With this display we adjust the line stretcher to flatten out the phase slope and then offset the phase display with the phase vernier to obtain an approximately 0° reading for



(a)



$$\text{Measured } s_{21} \equiv \frac{b_3}{a_0} = t_{30} + \frac{s_{21}\, t_{32}}{1 - s_{11}\, t_{11} - s_{22}\, t_{22} - s_{21}\, s_{12}\, t_{11}\, t_{22} + s_{11}\, t_{11}\, s_{22}\, t_{22}}$$

$t_{30} \simeq$ crosstalk

$t_{32} \simeq$ frequency tracking of power splitter & receiver

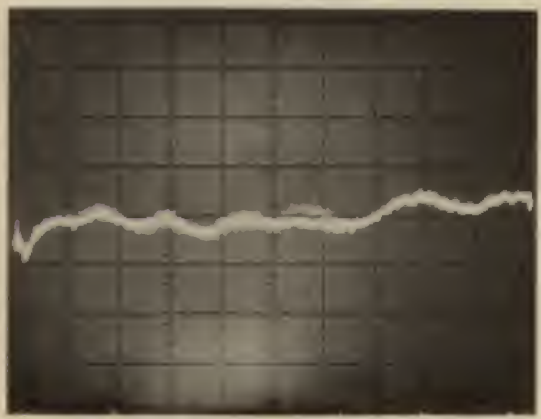
$t_{11} \simeq$ port 1 mismatch

$t_{22} \simeq$ port 2 mismatch

(b)

FIG 17-39 Transmission measurement system: (*a*) block diagram, and (*b*) error model.

(a)



(b)

FIG 17-40 Equalizing electrical length in transmission measurement system by using a phase-versus-frequency display: (*a*) swept measurement of through line before electrical length equalizing, and (*b*) result of equalization and adjusting phase vernier control.

all frequencies (see Fig. 17-40). If desired, the line stretcher can be used to remove or add electrical length to the network under test.

In making swept measurements, the frequency tracking error can be approximately isolated by connecting both ports together and drawing grid lines for each setting of the intermediate-frequency attenuator, in much the same way as illustrated in Fig. 17-36 for reflection measure-

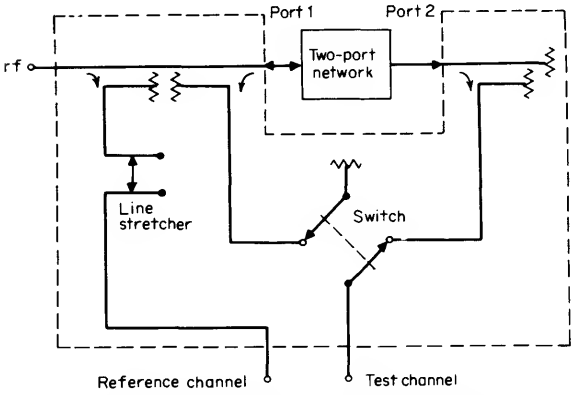


FIG 17-41 Combined reflection and transmission test set.

ments. The mismatch terms may be reduced by adding pads to port 1 and 2.

Incremental testing is a popular way to remove measurement system errors. Incremental testing compares two like networks, one which is well characterized and the other to be checked for deviation from the known network. As long as the unknown network does not deviate too far from the known network, the tracking and mismatch errors are comparatively small.

Multiparameter Measurements. Many times we should like to measure more than one s parameter with the same test set. Figure 17-41 shows a combined transmission and reflection test set. The switch position determines which parameter we are measuring. If any of the error terms in the transmission or reflection error models do not change very much with switch position, we may be able to combine the reflection and transmission error models as shown in Fig. 17-42. Here we have assumed that the port mismatch terms r_{11} and r_{22} do not change when the receiver is

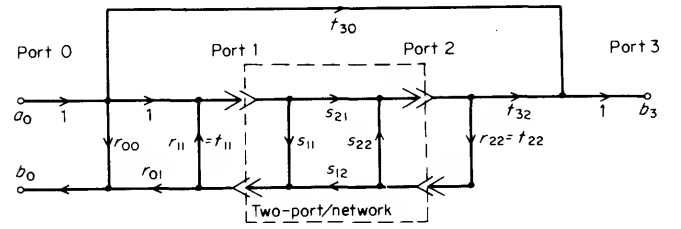


FIG 17-42 Error model for combined reflection and transmission measurement system.

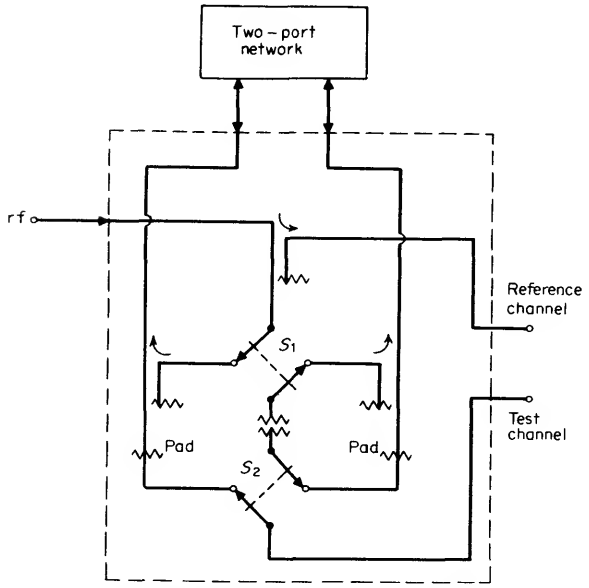


FIG 17-43 Four-parameter (s_{11}, s_{22}, s_{21}, and s_{12}) s-parameter test set.

switched from the reflection port to the transmission port. The analysis and equations for this test set and the operation of it are the same as described in the previous reflection and transmission measurement sections. If we had a three-channel receiver, we could eliminate the need for the switch and simultaneously monitor the transmission and reflection.

Figure 17-43 illustrates a four-parameter test set for measuring s_{11}, s_{21}, s_{22}, and s_{12}. With s_1 and s_2 set as shown in the diagram, the ratio of the test to reference channels is proportional to s_{11} of the network being tested. If s_1 and s_2 are both switched, we measure s_{22} of the network. If s_1 or s_2 are switched separately, we measure s_{12} or s_{21} of the network respectively.

Computer Measurement Systems. Highly sophisticated microwave devices and systems have created a need for accurate, fast, and complete characterization of the networks that comprise them. We have discussed two popular techniques for characterization of microwave networks. They can be broadly classified as fixed-frequency or swept-frequency techniques. The power of the fixed-frequency technique is that the mismatch, tracking, and directivity errors of the measurement system can be minimized by "tuning out" the residual errors and high accuracy can thereby be achieved. Fixed-frequency techniques, how-

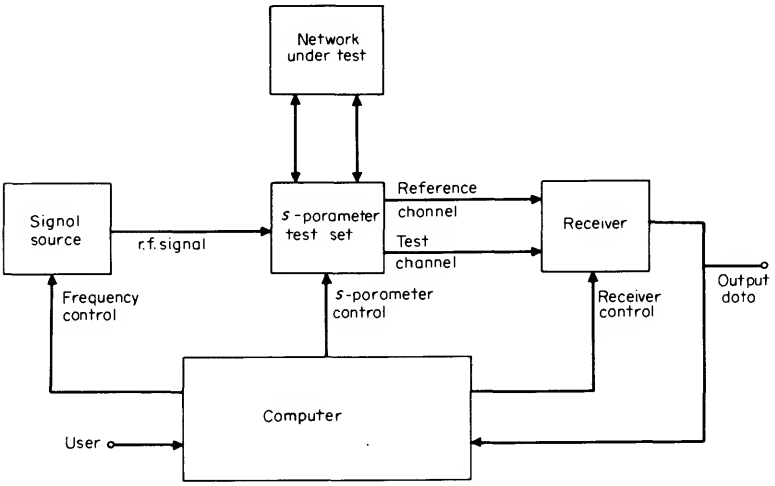


FIG 17-44 Computer-controlled automatic network analyzer system.

ever, are slow and somewhat tedious. Swept-frequency techniques in general offer a fast means of gathering data across broad bandwidths and the advantage of intuitive insight into the device being tested. Normally, it is difficult to account for all errors when using the swept-frequency technique, since they cannot be completely tuned out in the broad bandwidths. We should like the microwave measurement system to provide the advantages of the above two techniques without incurring the disadvantages and to be flexible and easy to use.

If we take the network analyzer system shown in Fig. 17-30 and add a computer, we obtain the automatic network analyzer shown in Fig. 17-44. The computer controls the various blocks and also processes the data from the receiver. Let us now discuss how the automatic system strengthens the following four areas.

Accuracy. The automatic system uses a stepped-frequency sweep rather than a continuous one, and so a finite number of points for measurement results. Instead of tuning out the system errors at each frequency point. the various error terms (see Fig. 17-32) are first measured, then taken into account as the device is measured. The internal system errors are vectorially subtracted from the measurement data, which corrects the measurement and leaves only the true characteristics of the device. System errors need only be characterized at the *beginning* of a set of measurements. The various error terms of the automatic network analyzer can be constructed by measuring appropriate standards. The standards

used are such devices as shorts, opens, and loads, which are relatively easy to characterize and manufacture to accurate tolerances.

The residual system errors, all of second-order importance, are then only those caused by imperfect repeatability of connectors and switches, noise in the system, system drift, and errors in the standards used for calibration.

The measurement procedure is as follows: (1) calibration, (2) measurement, and (3) correction of the data.

Speed. The computer can easily control all the instrument functions normally operated by the user. Then, too, the calculating power of the computer greatly shortens the time required for complete network characterization.

Flexibility. The *s* parameters are the parameters most easily measured at microwave frequencies. However, they may not be the desired output from the system. The *s* parameters comprise a total characterization of the network. Therefore, the computer can transform from the *s*-parameter set to any other consistent parameter set one may wish. The *h*, *y*, or *z* parameters can be determined, and also group delay, vswr, return loss, or other desired quantities. Transformations into other domains are also feasible, such as determining time-domain response from frequency-domain data.

Ease of Use. Since the computer controls the instruments, makes the measurements, and manipulates the data, the user is relieved of the mundane and difficult parts of the measurement procedure. The imagination of the development or production engineer is not merely supplemented; it is indeed amplified by the system, and furthermore his ideas, now in software, are made usable by many people. More time becomes available, and more desire is created to do the long and difficult measurements needed for imaginative design.

CITED REFERENCES

1. Adam, S. F.: "Microwave Theory and Applications," Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969.
2. Wholey, W. B., and W. N. Eldred: A New Type of Slotted Line Section, *Proc. IRE*, vol. 38, no. 3, March, 1950.
3. Sorger, G. U., and B. O. Weinschel: Swept Frequency Height Resolution VSWR Measuring System, *Weinschel Engineering Co., Internal Rept.* 90–117, p. 723, March, 1966.
4. Adam, S. F.: Swept VSWR Measurement in Coax, *Hewlett-Packard J.*, vol. 18, no. 4, December, 1966.
5. Ely, P. C.: Swept Frequency Techniques, *Proc. IEEE*, vol. 55, no. 6, pp. 991–1002, June, 1967.
6. Ely, P. C.: Swept Frequency Techniques, *Proc. IEEE*, vol. 55, no. 6, June, 1967.
7. Beatty, R. W.: Insertion Loss Concepts, *Proc. IEEE*, vol. 57, no. 6, June, 1969.

8. Fresh, D. L.: Methods of Preparation and Crystal Chemistry of Ferrites, *Proc. IRE*, vol. 44, pp. 1303–1311, October, 1956.
9. Lax, B. L.: Frequency Loss Characteristics of Microwave Ferrite Devices, *Proc. IRE*, vol. 44, pp. 1368–1386, October, 1956.
10. Heller, G. S.: Ferrites as Microwave Circuit Elements, *Proc. IRE*, vol. 44, p. 1389, October, 1956.
11. Weiss, M. T.: Improved Rectangular Waveguide Resonance Isolators, *IRE Trans. Microwave Theory Tech.*, vol. 4, pp. 240–243, October, 1956.
12. Heller, G. S., and G. W. Catuna: Measurement of Ferrite Isolation at 1,300 Mc, *IRE Trans. Microwave Theory Tech.*, vol. 6, p. 97, 1958.
13. Duncan, B. J., L. Swern, K. Komiyasu, and J. Hannwacker: Design Considerations for Broadband Coaxial Line Isolators, *Proc. IRE*, vol. 45, pp. 483–490, April, 1957.
14. Fleri, D., and G. Hanley, Nonreciprocity in Dielectric Loaded TEM Mode Transmission Lines, *IRE Trans. Microwave Theory Tech.*, vol. 9, pp. 23–27, January, 1959.
15. Weisbaum, S., and H. Seidel: The Field Displacement Isolator, *Bell System Tech. J.*, vol. 35, p. 877, 1956.
16. Weisbaum, S., and H. Boyet: Field Displacement Isolators at 4, 6, 11, and 24 KMC, *IRE Trans. Microwave Theory Tech.*, vol. 5, p. 194, 1957.
17. Treuhaft, M. A.: Network Properties of Circulators Based on the Scattering Concept, *Proc. IRE*, vol. 44, pp. 1394–1402, October, 1956.
18. Bosma, H.: On Stripline Y-circulation at UHF, *IEEE Trans. Microwave Theory Tech.*, vol. 12, pp. 61–72, January, 1964.
19. Fay, C. E., and R. L. Comstock: Operation of the Ferrite Junction Circulator, *IEEE, Trans. Microwave Theory Tech.*, vol. 13, pp. 15–27, January, 1965.
20. Simon, J. W.: Broadband Strip-transmission Line Y-junction Circulators, *IEEE Trans. Microwave Theory Tech.*, vol. 13, no. 3, pp. 335–345, May, 1965.
21. McChesney, G., and V. Dunn: Broadband, Lumped Element UHF Circulator, *IEEE Trans. Microwave Theory Tech.*, vol. 15, pp 198–199, March, 1967.
22. Ince, W. J., and E. Stern: Computer Analysis of Ferrite Digital Phase Shifters, *IEEE Intern. Conv. Record*, 5, 1966.
23. Reggia, F., and E. G. Spencer: A New Technique in Ferrite Phase Shifting for Beam Scanning of Microwave Antennas, *Proc. IRE*, vol. 45, p. 1510, 1957.
24. Kurokawa, K.: Power Waves and the Scattering Matrix, *IEEE Trans. Microwave Theory Tech.*, vol. 13, March, 1965.

CHAPTER EIGHTEEN

AUTOMATED MEASUREMENT SYSTEMS

From notes by

M. D. Ewy

Hewlett-Packard Company
Palo Alto, California

A measurement system is a combination of instruments working together to measure, sometimes to analyze, and to present the resultant data in a useful form.

An *automated* system is economically advantageous if:

1. Many repetitive measurements must be made or operations performed; or
2. A complicated operation requiring a high degree of skill must be performed.

Three types of automatic measurement systems will be described, which illustrate the automatic system capabilities of

1. Data acquisition
2. Data analysis
3. Programming and control or automatic testing

At the time of the present writing [1], "automatic testing of electronic

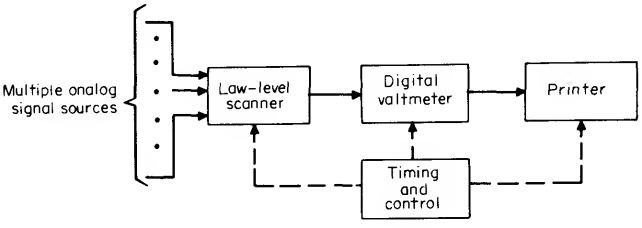


FIG 18-1 Basic low-level data acquisition system.

devices is one of the fastest growing areas of electronic instrumentation. It is the way an increasing amount of our testing is done every day, and it is the way most of our testing will be done in the future." Reference 1 contains four papers on automated testing, choosing an automatic test system, building an automatic test system, and a description of the systems produced by a leading company.

18-1 Low-level Multichannel Data Acquisition Systems

In contrast to the other automatic systems to be described, this type of system has been in use for many years. Applications have increased manyfold over the last two decades. The applications are often characterized by the need for monitoring, at many points, the condition of a process or equipment, for example, a nuclear reactor, aerodynamic structure, rocket, or jet or internal-combustion engine. In other words, the need is for automatic multichannel scanning; that is, each point to be monitored must be measured rapidly, one after the other, in sequence. It is often desirable to convert each measurement to digital form for outputting or recording to allow later accurate analysis.

A block diagram of a basic system is shown in Fig. 18-1. As the scanner successively connects the voltmeter input to each signal source, the voltage level from the source is measured and digitized by the voltmeter and presented in digital form to the printer for recording.

Frequently, the parameters to be measured by a low-level data acquisition system are temperature and pressure or force. These parameters are ordinarily converted into electrical signals by thermocouples for temperature measurement and strain gages for the measurement of pressure or force. A typical thermocouple generates a potential of 22 μV per °F. A strain gauge might generate 20 mV full scale. This situation requires the measurement of low-level potentials of the order of a few microvolts.

At this level, measurement capability is often limited by the errors introduced by common-mode and normal-mode noise signals. Signals

introduced identically on both the high and low input lines are referred to as *common-mode signals* (see Fig. 18-2). Figure 18-3 shows how both dc and ac common-mode noise signals are typically introduced into a data acquisition system. The desired signal source is the small differential

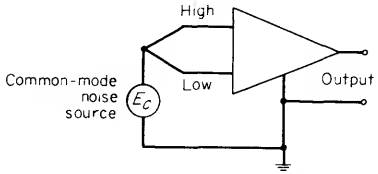


FIG 18-2 Common-mode voltage.

voltage developed across the resistance bridge through bridge unbalance. The excitation voltage for the bridge introduces a dc common-mode voltage, and a ground potential caused by ac power currents flowing in ground leads having finite impedance introduces an ac common-mode voltage. Both of these undesired voltages are often very large compared with the desired signal voltage. If the measurement system cannot reject these common-mode voltages well enough, then significant errors will appear at the output.

No amplifier or measurement system has perfect (zero) CMR ratio. Because of various unbalances in the source and in the measurement-system input circuits, some of the common-mode noise is converted into normal-mode noise. A normal-mode signal is a signal that appears between high and low, just as the desired signal appears. See Fig. 18-4 and also Chap. 13.

Several techniques are used in the design of low-level data acquisition systems to minimize the effects of noise. Most of these design techniques are described below. Also, see Chaps. 7 and 13 for discussion of the amplifiers themselves.

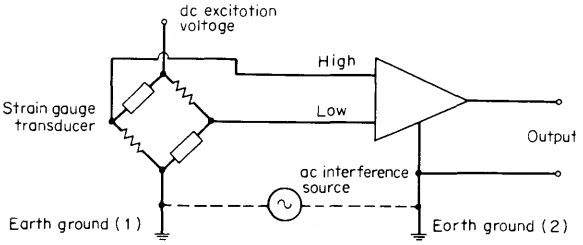


FIG 18-3 Alternating- and direct-current common-mode voltage.

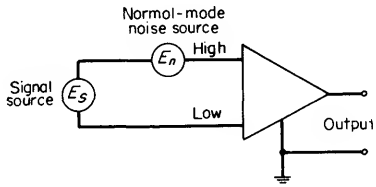


FIG 18-4 Normal-mode voltage.

18-2 Common-mode Noise Rejection

Floating. A floated measurement system has no dc connection between the chassis or earth ground and the measurement circuits; that is, the measurement circuits are floated. Figure 18-5 shows the significant parts of a bridge transducer with its output measured by a floated voltmeter. Reasonable circuit values are shown.

If the voltmeter were not floated, Lo would be connected to chassis ground and essentially all the voltage generated by the noise generator would appear across R_L and thereby interfere with the signal being measured. Floating reduces the interference to the extent that $X_{C1} = 1/\omega C$ is large compared with R_L. For the values shown, CMR ratio is approximately

$$\text{CMR (dB)} \approx -20 \log \frac{X_{C1}}{R_L} = -20 \log \frac{1}{\omega C_1 R_L} \qquad (18\text{-}2\text{-}1)$$

$$\approx -48 \text{ dB}$$

For long input lines, the use of shielding as shown does not improve the noise rejection because of the low value of R_C. An insignificant part of the current generated by the noise generator normally flows through R_H because of the high input impedance required of the voltmeter. For

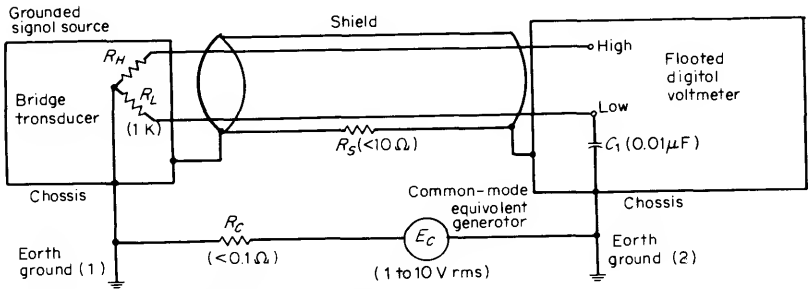


FIG 18-5 A floated measurement system.

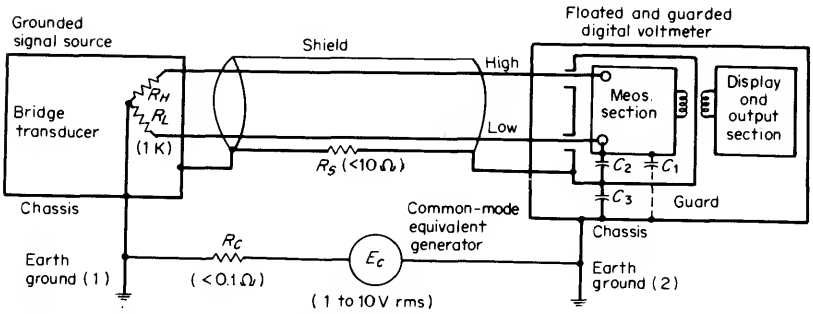


FIG 18-6 A floated and guarded measurement system.

some special applications, it would be possible to improve the situation further by maintaining appropriate impedances in the Hi and Lo lines at the source and the voltmeter input circuitry.

Guarding. A guarded measurement system has an electrostatic shield, or guard, surrounding the input lines and measurement circuits (see Fig. 18-6). The common-mode noise generator now drives the guard by charging C_3 through R_S. Since the potential on the guard shield nearly follows the potential on the Lo input line, very little noise current flows through R_L and C_2. Also, the guard shield can be made quite effective, so that C_1 is only a few picofarads. Therefore, very little noise current flows directly through R_L and C_1.

Assuming reasonable values of 0.01 μF for C_2 and C_3 (the same value used for C_1 in Fig. 18-5), 3 pF for C_3, and 60 Hz,

$$X_{C1} \gg R_L$$
$$X_{C2} \gg R_L$$
$$X_{C3} \gg R_S$$

The noise voltage developed across R_L is approximately

$$E_n \approx \left(\frac{R_S R_L}{X_{C2} X_{C3}} + \frac{R_L}{X_{C1}}\right) E_C$$

$$\text{CMR (dB)} = -20 \log \frac{V_C}{V_n}$$

$$\approx 20 \log \left(\frac{R_S R_L}{X_{C2} X_{C3}} + \frac{R_L}{X_{C1}}\right)$$

For the parameters given,

$$\frac{R_S R_L}{X_{C2} X_{C3}} \ll \frac{R_L}{X_{C1}}$$

and

$$\text{CMR} \approx -20 \log \frac{X_{C1}}{R_L} = -20 \log \frac{1}{\omega C_1 R_L} \quad \text{decibels} \qquad (18\text{-}2\text{-}2)$$

Note that this formula is identical with (18-2-1), but in this case the value of X_{C1} is orders of magnitude larger. For the values given, CMR ≈ -119 dB.

If leakage resistances R_1, R_2, and R_3 are substituted for the capacitors C_1, C_2, and C_3 of Fig. 18-6, it can be seen that interference from a dc common-mode voltage will be similarly rejected by a floated and guarded measurement system. The unguarded leakage resistance R_3 can be as high as 10^{11} Ω in a practical instrument, which gives −160-dB CMR at dc for 10^3-Ω source unbalance.

As indicated in Fig. 18-6, the measurement information must be coupled through the guard shield with ac coupling to preserve the integrity of the shield. Low-level DVMs have measurement techniques that generate pulses, which are ac coupled through the guard shield and then counted or totaled in output circuitry that can be referenced to chassis or earth ground [2, 3].

18-3 Normal-mode Noise Rejection

In the preceding section, we have shown how floating and guarding can reject a 60-Hz common-mode signal by a ratio of 10^6 or 120 dB. If a system experienced 10 V rms of common-mode signal because of power-line ground currents, only 10 μV rms would be converted to normal-mode interference and appear in series with the desired signal. However, even an error of 10 μV is large for systems measuring the outputs of strain gages or thermocouples. The parameters measured with strain gages and thermocouples usually fluctuate only at very low frequencies. Therefore, the interfering 60-Hz normal-mode signal can often be attenuated by a low-pass filter of some kind.

Other noise, including amplifier noise (see next section), is also significant at the 1-μV level. Therefore, a desirable filter not only would have high attenuation at 60 Hz, but also good attenuation at all higher frequencies. The filter should have minimum attenuation in the passband. Because the DVM is preceded by a scanner, a filter at the input terminals of the voltmeter must have fast transient response. Otherwise, accurate measurements could be made only at slow scanning rates, to allow the filter output to settle, or a preamplifier and filter would have to be used on every channel ahead of the scanner.

A technique often used, which meets all of the requirements above, is

electronic integration over a period of time equal to one or several periods of the interfering 60-Hz signal [2, 3]. (See Chap. 7.)

Note that even though the equivalent noise bandwidth is approximately 30 Hz with the integration period set at $\frac{1}{60}$ sec, a preamplifier circuit ahead of the integrating amplifier can be wideband and settle quickly after each scanner connection.

18-4 Low-noise Preamplifiers

As previously shown, it is often desirable to measure or resolve 1 μV, or even 0.1 μV, with a source impedance of 10^3 Ω in order of magnitude. An integrating DVM with an integration period of $\frac{1}{60}$ sec has an equivalent noise bandwidth of approximately 30 Hz. Thermal noise in a pure resistance is given by

$$E_n = (4KTR\,\Delta f)^{1/2} \tag{18-4-1}$$

where E_n = noise voltage, V rms
K = Boltzmann's constant, 1.38×10^{-23} j/°K
T = absolute temperature, °K
R = resistance, Ω
Δf = equivalent noise bandwidth, Hz

At room temperature, and with the values given above, an ideal source resistance of 1,000 Ω would generate a noise voltage of 0.022 μV.

A preamplifier not carefully designed for low-noise performance might have a noise figure of 20 dB, especially at these low frequencies. The equivalent input noise of this amplifier would be, for the 10^3 Ω source, 0.22 μV. Design techniques for obtaining lower noise figures are described in many books and papers on amplifiers.

18-5 Crosstalk

Crosstalk refers to the normal-mode interference on a channel being measured, generated by the signal or signals on other channels. The interference present in a scanning system is due to leakage impedance between signal lines and across open scanning switches. Figure 18-7 shows the basic situation, where

E_{s1} = signal being measured
E_{sn} = interference signal
R_1, R_n = source resistances
R_p, R_s = leakage resistances
C_p, C_s = leakage capacitances
R_i = input impedance of measuring circuit

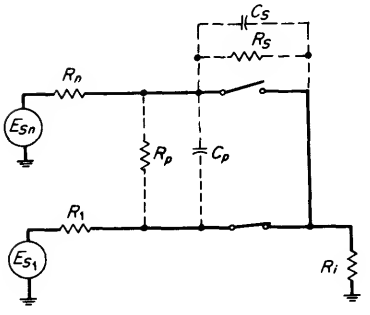


FIG 18-7 Crosstalk parameters (only Hi-line switches are shown).

The total leakage impedance is

$$Z_t = \frac{R_s R_p}{R_s R_p \omega(C_s + C_p) + R_s + R_p} \tag{18-5-1}$$

At dc,

$$Z_t = R_t = \frac{R_s R_p}{R_s + R_p} \tag{18-5-2}$$

and at high frequencies,

$$Z_t = X_t = \frac{1}{\omega(C_s + C_p)} \tag{18-5-3}$$

Crosstalk is normally not a problem at frequencies approaching dc. In typical applications, $R_i \gg R_1$ and at frequencies of interest $X_t \gg R_n$ and R_1. Then the crosstalk ratio CT is

$$\text{CT} = R_1 \omega(C_s + C_p) \tag{18-5-4}$$

or

$$\text{CT (dB)} = 20 \log R_1 \omega(C_s + C_p) \tag{18-5-5}$$

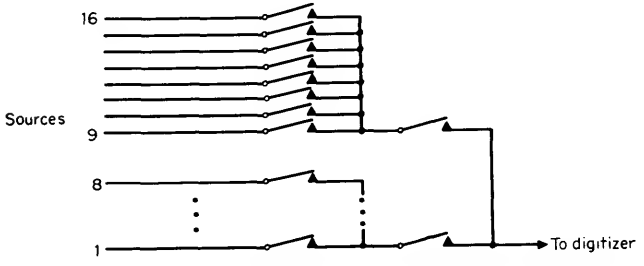


FIG 18-8 Two-level switching to minimize crosstalk (only Hi-line switches are shown).

In most applications, C_p can be reduced by electrostatic shielding or guarding, but C_s is additive for each source or channel connected to the system. Figure 18-8 shows how two-level switching can be used to minimize the effective C_s in a scanning system.

18-6 Thermal Voltages

A junction of two dissimilar metals will generate a small voltage, or emf, with amplitude depending upon temperature. This result is commonly called a *thermal emf*, since the magnitude of the voltage is

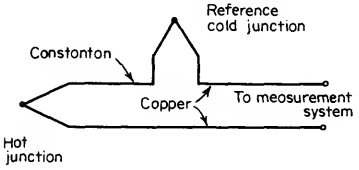


FIG 18-9 Thermocouple.

a function of temperature. Thermocouples are junctions of dissimilar metals fabricated specifically for measuring temperature by using this effect, and a diagram of such a device is shown in Fig. 18-9. When both junctions are at the same temperature, the thermal emf's at the junction are of equal magnitude and opposite polarity, so that the net voltage seen by the measurement system is zero. If the temperatures of the two junctions are different, then a net voltage is developed that is proportional to this differential temperature. The metals shown will develop a net voltage of about 22 μV/°F of temperature difference between hot and cold junctions.

This effect is useful, but great care must be taken to prevent undesired thermal emf's from existing in the low-level input circuits. An input

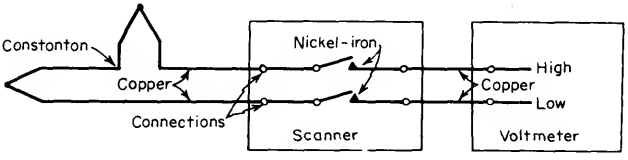


FIG 18-10 Thermocouple input circuit.

circuit is shown in Fig. 18-10 that uses a copper-constantan thermocouple and a magnetic-reed scanner to route the signal to a low-level voltmeter. For clarity, the guard circuit that is normally used is not shown. There are typically many more connections than shown here.

Now, to measure a temperature to an accuracy of 0.1°F would require measurement of the input voltage to an accuracy of about 2 μV. The thermal emf developed by the reed scanner, if there is a temperature difference at the two junctions of nickel-iron to copper, is greater than 10 μV/°F. Careful thermal design must be used to ensure that the temperature difference between the two ends of the reed switch is held very low. Also, if care is not taken, the connecting lines may contain materials having different thermoelectric coefficients, and unless junction pairs are accurately matched in temperature, a residual error voltage will exist and drift about with changes in temperature. For instance, if copper wires and brass connectors are used, a temperature change of 1°F at one of the junctions will cause a drift of more than 1 μV referred to the input.

In some cases, even the different thermoelectric coefficients of wire from different ingots of copper can cause significant error. And if two pieces of copper wire are soldered together without clean copper-to-copper contact, *two thermoelectric junctions exist at the connection* and can cause drift.

18-7 Scanners

Low-speed Scanners. The techniques described above allow the design of systems that provide accurate measurements down to the microvolt level in very difficult environments. However, the integration technique for eliminating 60-Hz normal-mode interference limits the system speed to a maximum of about 40 measurements per second. At these speeds, magnetic-reed switches and other electromechanical switches can be used to perform the input scanning function. These scanners switch three wires, HIGH, LOW, and GUARD, for each input channel, and connect these wires to a cable routed to the DVM. The switching typically requires a few milliseconds.

High-speed Scanners or Multiplexers. A high-speed scanner is usually called a *multiplexer*. The electromechanical scanner is not fast enough to be used in high-speed data acquisition systems. Solid-state switches are used instead. These switches must be followed by a wideband, low-noise amplifier with low output impedance, that is physically close to the output of the switches. This allows higher system speeds by minimizing the capacitance on the output side of the switch that must be driven by the source. Suppose, for example, that a data acquisition system uses a 10-ft signal cable, as shown in Fig. 18-11. With a reasonable cable capacitance of 30 pF/ft, C_i = 300 pF. Further, assume that the voltage on C_i is 0 V before the scanner is connected to E_{s1} and that $R_1 = 10^5$ Ω. If a reading to an accuracy of 0.1 percent is desired, R_i must be

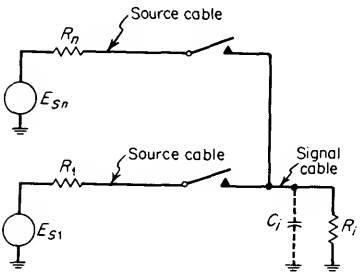


FIG 18-11 Input capacitance effect (only Hi-line switches are shown).

$> 10^8\ \Omega$ to prevent steady-state loading effects, and this resistance is easy to obtain. However, one must wait more than 6.9 time constants τ, where $\tau = R_1C_i$, before making a measurement, to allow the voltage on C_i to rise to within 0.1 percent of E_{s1}. Then this time is $T > 6.9\tau = 6.9 \times 10^5 \times 300 \times 10^{-12}$, or $T > 207\ \mu$sec.

This time is insignificant if one is integrating the input signal over a period of $\frac{1}{60}$ sec or longer, as in low-speed systems. However, it is seriously long in high-speed systems. These systems are often required to make measurements to 0.1 percent accuracy in 10 μsec or less. Therefore, a high-performance amplifier must be used near, and on the output side of, the switches.

Other techniques to minimize the effective value of C_i are two-level switching (see Fig. 18-8) and the use of multiple amplifiers, with only a limited number of sources switched into any one amplifier.

Analog-to-digital Conversion Techniques. In many modern data acquisition systems, the scanner or multiplexer is followed by an instrument that can measure the input analog voltage and present the resultant number in digital form for display or recording, or both. In slow-speed systems, this device is usually a DVM. The integration technique previously referred to is often used because it provides outstanding noise rejection capability.

In high-speed systems, the instruments that perform the measurement and conversion are called *analog-to-digital* (or A/D) *converters*. Some methods for analog-to-digital conversion are treated in Chap. 8.

18-8 Automatic Analyzer Systems

In an application suited to the use of an automatic analyzer system, the measurement requirement is not necessarily very repetitive, but considerable analysis must be made to obtain the desired answer. Small computers are therefore appropriate in this type of system. An example is the microwave-network analyzer system shown in block diagram form in Figure 18-12. See Ref. 8 for details.

The signal source consists of a sweep signal generator interfaced to the computer for the control of generator frequency.

The network analyzer measures the s parameters [4] of the device being characterized. Any one of the four s parameters representing a two-port device can be measured by switching the rf signals inside the test unit. This switching is computer controlled. The network analyzer portion of the system is a microwave vector-ratio voltmeter which measures the amplitude ratio and phase difference between the reference and test channels [5, 6]. Thus, for the position of the test-unit switches shown in Fig. 18-12, the port-1 reflection characteristics of the device under test would be measured. Two analog-to-digital converters in the network analyzer digitize the magnitude and phase information for computer input.

The computer is a small instrumentation computer [7]. All important functions of the microwave instruments can be controlled by the computer. Also, the computer can be used to store the system's own frequency-dependent errors (both amplitude and phase) as well as perform necessary mathematical calculations. Control of the system is as follows: First the computer sets the sweep generator and network analyzer to the required frequency and then routes the rf signal through the multiplexer to the test unit. Next, the s parameter to be measured is selected, and the resulting amplitude and phase information is digitized and trans-

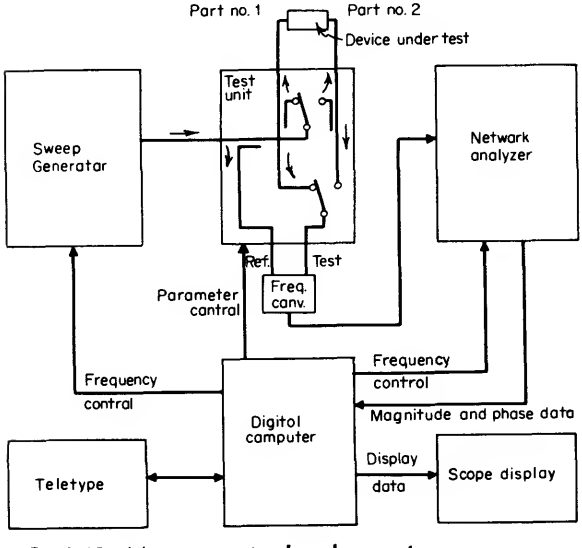


FIG 18-12 Microwave-network analyzer system.

ferred to the computer. Finally, the raw measurement data are corrected for system errors (previously measured) and then transformed into the desired parameters for display on either an oscilloscope or the teletype. These last two operations of the computer (error correction and data transformation) are very important. They are the unique characteristics of this kind of system and make it very useful.

Error Correction. The error correction capability allows the system to make microwave measurements to the accuracy of a standards laboratory quickly and easily. This is accomplished by making s parameter measurements with reference standards connected to the measurement ports, and also with the ports connected together. Then simultaneous equations are solved and other calculations made to derive eight separate calibration constants. This calibration procedure is automatically performed at each frequency of interest. Then, after the s parameters on the unknown device are measured, they are corrected by an iterative computing technique that uses the stored calibration data, and this yields measurements with standards laboratory accuracy in only a few minutes.

Data Transformation. The data transformation capability of the computer comes into play after error correction. The corrected s-parameter measurements have completely characterized the device, but the user normally wants the information in the form of VSWR, reflection coefficient, return loss, bandpass transmission, or group delay, all these versus frequency. The digital computer can quickly calculate any or all of these characteristics and print them out on a teletype or plot them on a cathode-ray oscilloscope display.

Thus it is seen that small digital computers make it possible to build automatic analyzer systems that make accurate measurements quickly, and these would otherwise be made inaccurately or not at all because of the tedium involved.

18-9 *Automatic Test Systems*

An automatic test system application typically requires not only some data acquisition and analysis, but also a considerable amount of programming and control capability. This kind of system is often used to check the performance of some device on a production line or at a maintenance facility. The applications vary from electronic component testing such as capacitor testing, to large system testing such as the checkout of all electronic systems on an aircraft. We shall describe here a typical test system that would be useful for testing electronic assemblies, such as printed circuit cards or a receiver. A basic block diagram is shown in Fig. 18-13.

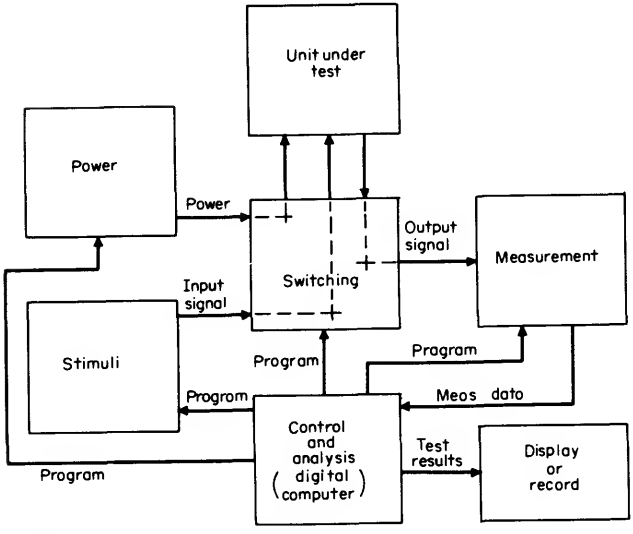


FIG 18-13 Automatic test system.

The system would operate basically as follows: After the unit under test is connected to the system, the switching unit would be programmed to make the proper system interconnections, and then the power would be programmed ON to the proper level. The stimuli are then programmed ON, and after an appropriate time interval, a measurement might be compared with a desired result, and if the comparison is satisfactory, the system would be programmed for the next test to be made. At the completion of all tests on the unit, the significant results are recorded on an output device. The entire sequence of tests can be made automatically, without operator intervention. Almost any measurement that can be made manually on a device can now be made automatically with such a system, with a great saving of time and with increased repeatability and accuracy.

A system might include several power supplies, stimuli, and measurement instruments. The stimulus sources could include, for example, low- and high-frequency oscillators, function generators, dc sources, and pulse generators. The measurement instruments could include DVMs for measuring low-level or dc signals, analog-to-digital converters for making high-speed measurements of amplitude versus time, ac DVMs and phase meters, digital ohmmeters, and counters for measuring frequency or time intervals. The system could also include programmable loads, attenuators, amplifiers, and calibration references [8].

With instruments such as these, the system could be used automatically to measure the gain and phase versus frequency, the time response to a step input, and time delays and logical operation of digital circuitry, for example. The measurements made are usually the same measurements that could be made manually, but the computerized system can dramatically speed up the process. The computer can also be used to analyze the results more thoroughly and even to perform diagnostic troubleshooting on test units that exhibit improper performance.

The instruments used in such systems are in most respects the same as the instruments used in manual test setups. The primary difference is that the instruments must be programmable. That is, the range-changing, level-setting, switching, etc., that would otherwise be done manually must be programmed by the computer. Also, data generated by the measurement instruments must be in a form suitable for reading by a computer.

Programming. Digital computers are used for control and analysis in modern automatic test systems [7]. All information in such a computer is in the form of digital bit patterns. The computer normally operates on (that is, stores, transfers, reads, or performs logical arithmetic operations on) a bit pattern of a fixed length, called a *character* or *word*, one word at a time. This is then the natural means of communicating between the digital computer and other instruments in the system. That is, the computer would like to transfer data to or read data from an instrument one word at a time, each word containing, for example, 16 bits of digital information. The instruments then need to recognize this digital programming data. This requirement is easily met for those instrument adjustments that would be made manually by switching—for example, range changing. A switch connection can be remotely controlled by using a relay or a transistor that is biased OFF or saturated ON in the programmed instrument. The first eight bits in the program word to an oscillator, for example, could control the switching of eight frequency ranges. Each bit is represented by an electrical signal, and only one of the eight is caused to be ON or TRUE at any one time and thereby select the appropriate range. A further refinement could binary encode this range program data, which would require the transfer of only three data bits from the computer to the instrument.

Sometimes program data take the form of numbers, for example, the channel number to which a scanner should be connected. This information is transferred as a binary-coded octal or binary-coded decimal number. Some instrument controls, such as signal output-level controls, are analog in nature. Here a digital-to-analog converter is needed to transform the digital program data before it is used to program the instrument. The digital-to-analog converter could be a part of the pro-

grammed instrument, or it could take the form of separate interface hardware.

Addressing and Memory. All digital program data transferred out of a digital computer are transferred from a common data buss in the computer. The program data must somehow be directed to the appropriate instrument. This is done by addressing. Each instrument has a unique address, an octal number. When an instrument is addressed, the computer output data buss is connected to that instrument, or to a register associated with that instrument. This register then remembers the data that were transferred. The data buss then can be used to transfer data to another instrument. This register, or memory, is essential. It must exist to hold the program data as long as the program is needed. The registers and associated addressable switches could be in the instruments, in an interface package between the instruments and the computer, or within the computer main frame.

Timing. A typical modern small digital computer contains a clock that causes it to operate on the order of 10^6 operations per second [7]. Such a computer could successively read program data from its memory and transmit the data to instruments to connect system switches, program power supplies, program an ac source, program an ac voltmeter, and examine the voltmeter for measurement data, all in a time interval of less than 100 μsec. The instruments obviously cannot respond this fast. Typically, the switches would require a few milliseconds to close and stop bouncing, a fast power supply would require 100 μsec to settle, and an ac source and voltmeter could each take a good part of a second to settle and obtain an accurate measurement. The point is, system timing information must be incorporated into the computer program. There are basically two methods.

The most obvious approach is to program the computer to wait, or "count time," for an appropriate period after each time that it programs or communicates with an instrument. This is a completely workable approach, but it requires that the person programming the system (writing the computer program) have a good knowledge of the timing requirements of all instruments in the system. Another approach is to provide a line between the computer and the instrument, over which timing information is transferred. With such a line, an instrument can "tell" the computer when it has "timed out" or settled, by changing the digital level on this line at that time. The computer then simply monitors this line until the appropriate level change occurs and then proceeds with the next system operation. The system then operates at maximum permissible speed, determined by the instruments, without the need for putting explicit instrument timing information in the computer program.

More highly developed automatic test systems use this timing inform

tion from the instruments in an interrupt mode. The computer program can be so arranged that whenever the computer must wait for an instrument to settle or generate a reading (for example), the computer program goes on to other problems that can be done while waiting; such as performing calculations on data previously obtained. Then when the instrument interrupts at the end of the required instrument delay time, the computer program goes back to the operation with the instruments. This technique of course makes more efficient use of computer time.

CITED REFERENCES

1. *Hewlett-Packard J.*, vol. 20, no. 12, August, 1969.
2. Andersen, R. A.: A New Digital Voltmeter Having High Rejection of Hum and Noise, *Hewlett-Packard J.*, vol. 13, no. 6, February, 1962.
3. McCullough, William: A Fast-reading Digital Voltmeter with 0.005% Accuracy and Integrating Capability, *Hewlett-Packard J.*, vol. 16, no. 12, August, 1965.
4. Anderson, R. W.: S-parameter Techniques for Faster, More Accurate Network Design, *Hewlett-Packard J.*, vol. 18, no. 6, February, 1967.
5. Anderson, R. W., and O. T. Dennison: An Advanced New Network Analyzer for Sweep-measuring Amplitude and Phase from 0.1 to 12.4 GHz, *Hewlett-Packard J.*, vol. 18, no. 6, February, 1967.
6. Network Analysis at Microwave Frequencies, *Hewlett-Packard Company Appl. Note* 92, May, 1968.
7. Magleby, K. B.: A Computer for Instrumentation Systems, *Hewlett-Packard J.*, vol. 18, no. 7, March, 1967.
8. Hackborn, R. A.: An Automatic Network Analyzer System, *Microwave J.*, vol. 11, no. 5, May, 1968.

INDEX

INDEX

Absorption frequency meter, 615
Absorption modulator, 585
Ac (*see* Alternating current)
Admittance, 265
AGC (automatic gain control), 523
Alternating-current comparator, 282
Alternating-current probe, 261
Alternating-current resistance, 265
Amplifier:
  audio, 480
  chopper, 204
  common-mode rejection, 491
  deflection, 381
  differential, 201, 376, 496
  direct-coupled, 188
  direct-current, 186
  distortion, 500
  dynamic range, 500
  feedback, 485
  gain measurement, 484
  impedance, 482
  input balance, 491
  loop gain, 483
  low-noise, 710
  measurements, 480
  microwave, 552
  phase measurement, 48
  slew limiting, 505
  solid-state, 550
  transistor, 554
  tunnel-diode, 551, 553
Amplifier (*Cont.*):
  video, 480
Amplitude modulator, 584
Analog-to-digital conversion, 86
Analyzer:
  spectrum, 616
  wave, 632
Analyzer system, automatic, 714
Anhysteretic magnetization, 469
Aristarchus of Samos, 2
Astronomical observations, 2
Astronomy, 4
Atomic frequency standards, 170
Atomic time, 158
Automatic analyzer system, 714
Automatic gain control (AGC), 523
Automatic test system, 716
Avalanche oscillator, 569

Background disturbance, 102
Barretter, 604
Beat-frequency oscillator, 340
  polyphase, 342
Bode plot, 486
Brahe, Tycho, 2
Bridge:
  active, 301
  automatic, 302
  basic circuits, 292
  frequency, 291

Bridge (*Cont.*):
inductively coupled ratio arm, 298
Kelvin double, 285
loading error, 294
low-frequency, 288
megohm, 287
radio frequency, 304
special purpose, 297
Warshawsky, 286
Wheatstone, 282

Capacitance meter, 279
Capacitor:
frequency effects, 272
standard, 274
Cathode-ray tube (CRT), 353
graticules, 364
mesh expansion, 359
phosphors, 360
postacceleration, 357
storage-target, 365
variable persistence, 370
writing speed, 363
CCIT method, 504
Chebyshev line shapes, 132
Chebyshev window, 129
Chopper:
amplifier, 204
transistor, 207
Circulator, 666
Coherent detection, 99
Comité Consultatif International Télégraphique method, 504
Comité International des Poids et Mesures, 12
Common-mode rejection, 377, 491, 707
in digital voltmeter, 233
Communication system interferences, 536
Complex impedance meter, 280
Computer information display, 415
Conducted interference, 511
Conducted susceptibility, 515
Confidence level, 124
Convolution, 56, 113
definition of, 108
digital, 139
Cooley-Tukey algorithm, 110
Copernicus, 2
Correlation function, 76, 89
Crookes, William, 350
Crosstalk, 710
CRT (cathode-ray tube), 353
Cumulative probability distribution, 74, 113
Curve tracer, 416

Data acquisition system, 705
Dc (*see* Direct current)
Defense Communications Agency, 543
Delay distortion, 30
Delayed sweep, 387
DeMoivre-Laplace theorem, 116
Detection:
coherent, 99
of signal in noise, 100
Detector:
average-responding, 239
peak, 245
peak-to-peak, 249
phase, 25
quasi-root-mean-square, 253
root-mean-square, 250
Differential amplifier, 201
Differential time delay, 528
Digital convolution, 139
Digital filtering, 139
Digital voltmeter (DVM), 211
counting, 219
integrating, 82, 224
noise rejection, 230
potentiometric, 213
ramp, 218
Digitizing errors, 119
Dimensions, 10
Direct-coupled amplifier, 188
chopper, 204
differential, 201
open-loop gain measurement, 195
overvoltage protection, 194
Direct-current amplifier, 186
automatic reset, 197
Direct-current comparator, 281
Direct-current meter, 276
Direct-current probe, 259
Direct-current resistance, 265
Distortion, 643
delay, 30
measures, 41

Distribution, cumulative probability, 74, 113
Disturbance, background, 102

Electromagnetic susceptibility, 514
Emu, 7
Ensemble of noise signal, 66
Envelope delay, 31, 530
Ephemeris time, 157
Equivalent circuit, 266
Equivalent filter, 117
  for signal averaging, 118
Ergodic systems, 68
Esu, 7

Ferrimagnetic resonance, 662
Ferrite devices, 661
Filtering, digital, 139
Fourier transform, 47, 49
  basic theory, 108
  common pairs, 109
  Cooley-Tukey algorithm, 111
Frequency:
  broadcasts, 158
  ratio measurement, 176
  VLF broadcasts, 160
Frequency measurement, 177
  standard deviations, 164
Frequency measuring instruments, 172
Frequency meter:
  absorption, 615
  microwave, 611
Frequency multiplier, 556
  stability, 564
Frequency response, automatic plotting, 22
Frequency stability, 642
Frequency standards, 162
  atomic, 170
  quartz, 167
Frequency synthesizer, 179
  applications of, 183
Function generator, 348

Gain margin, 36
Gain measurement, 19, 489
Galvanometric recorder, 428
Group delay, 31, 34, 530
Guarded measurement, 708
Guarding, 708
Gunn oscillator, 565
Gyrator, 669

Hanning line shapes, 129
Hanning window, 129
Harmonic mixing, 630
Hewlett-Packard *RX* meter, 305

Impedance, 265
  precision measurement, 311
  transfer, 268
Impedance measurements, 264, 655
Impedance standards, 313
Impulse noise, 645
Impulse response, 47
  by crosscorrelation, 101
Inductance meter, 279
Inductance standards, 13
Inductively coupled frequency, 291
Inductor:
  frequency effects, 275
  standard, 276
Input balance, 491
Instruments, frequency measuring, 172
Integration, dual-slope, in digital voltmeter, 227
Interference, 536
  conducted, 511
  radiated, 514
Interference measurement, 653
International Committee on Weights and Measures (Comité International des Poids et Mesures), 12
International system of units, 12
Isolator, 664

Kelvin double bridge, 285
Kepler, Johannes, 2
Kerr bridge, 305

LC oscillator, 321
Line shape, 126, 127
Linear systems, analysis of, 17
Loop gain, 35

Loss measurement, 19
Low-noise amplifier, 710

Magnetic recording, 461
  analog, 468
  noise in, 475
  reproduction of, 472
Magnetic recording methods, 464
Magnetic tape, 476
Matched filter, digital, 141
Mathematics, 66
Mean solar time, 158
Mean-square values, 69
Measurement:
  automated systems, 704
  computer systems method, 701
  electromagnetic interference, 653
  guarding, 708
  impedance, 264, 655
  interference, 653
  low-impedance, 299
  microwave attenuation, 659
  microwave techniques, 688
  modulation, 639
  multiparameter, 699
  mutual-inductance, 300
  presence of noise, 97
  radio-frequency, 303
  reflection, 655, 689
  transmission, 695
Microwave amplifier, 552
Microwave attenuation measurement, 659
Microwave power measurement, 595
Microwave sweep generator, 588
Microwave transistor oscillator, 545
Microwave wavemeter, 611
Modulation, pulse amplitude, 647
Modulation measurement, 639
Modulator:
  absorption, 585
  amplitude, 584
Multiparameter measurement, 699
Multiplexer, 713
Multiplier:
  circuit, 558
  frequency, 556
  step-recovery diode, 562
  varactor-diode, 556

National Bureau of Standards (NBS), 159, 312
  station WWVB, 160
  traceability, 313
Network, electrical, 305
  bridged T, 306
Network analyzer, 29
  s-parameters, 686
Network theory, two-port, 671
Nichols plot, 487
Noise:
  background, 90
  common-mode rejection, 377, 491, 707
  effect of, on measurement of power spectrum, 138
  ensemble average, 66
  mean-square values, 69
  normal-mode rejection, 709
  power density spectrum, 70
  power spectrum, 69
  in recorders, 475
  spectral measurements, 87
  in spectrum analyzers, 628
  as a test signal, 87
  time average, 67
  white, 71
Noise figure, modulation, 521
Noise measurement, 80
Nonlinearity, classes of, 46
Normal-mode rejection, 709

Ohmmeter, 277
Optical recorder, 448
Oscillator:
  avalanche, 569
  beat-frequency, 340
  Gunn, 565
  impatt diode, 567
  LC, 321
  microwave transistor, 545
  phase-shift, 337
  polyphase, 342
  RC, 328
  ring, 339
  solid-state, 565
  stability, 578
  tuning, 548
  Wien bridge, 331
Oscilloscope:
  accessories for, 418

Oscilloscope (*Cont.*):
amplifier, 374
classification of, 352
delay lines, 392
digital readout, 407
functions of, 351
general-purpose, 371
for medical displays, 413
plug-in, 395
sampling, 398
storage, 397
sweep linearity, 388
sweep mode, 386
time-bases, 382
x-y plotter, 411
Oscilloscope camera, 421

PAM, 647
Period measurement, 175
Phase detector, 25
Phase margin, 36
Phase measurement, 23
Phase-shift oscillator, 337
Physical constants, table, 14
Polyphase oscillator, 342
Power density spectrum, 70
Power measurement:
microwave, 595
mismatch, 607
pulsed-power, 605
Power meter:
thermistor, 600
thermocouple, 598
Power sensor, 596
Power sensor mounts, 610
Power spectral analysis, 134
Power spectrum, 69
Precautions in testing, 42
Probability density function, 71, 112
Gaussian, 72
of sampling error, 115
Probe:
alternating-current, 261
direct-current, 259
Programming, 718
Pulse-amplitude modulation (PAM), 647
Pulsed-power measurement, 605

Q meter, 310
Quartz frequency standards, 167

Radiated interference, 514
Radio-frequency bridges, 304
Radio-frequency measurements, 303
Radio receiver, amplitude-modulation, 508
Random variable, 112, 113
Ratio measurement, frequency, 176
RC oscillator, 328
Reactance, 265
Receiver, 507
amplitude-modulation, 508
automatic-gain-control, 523
commercial specifications for, 539
federal and military specifications for, 540
modulation, 521
noise figure, 520
quieting, 516
Recorder:
amplifier for, 440
galvanometric, 428
noise, 475
optical, 448
pen mechanisms, 446
phase response, 438
position feedback, 454
servo, 456
sine response, 436
straight-line writing, 450
transient response, 430, 459
Recording:
digital, 466
magnetic, 461
Reflection measurement, 655, 689
Reflectometer, 657
Resistor:
frequency effects, 270
power coefficient, 270
standard, 272
temperature coefficient, 270
voltage coefficient, 269
Resonance method (impedance measurement), 308
Ring oscillator, 338
Root-mean-square (rms) detector, 250
Root-mean-square (rms) voltmeter, 250

Sampling oscilloscope, 398
Sampling voltmeter, 255
Scan conversion, 371
Scanner, 713
Scattering parameters, 673, 678, 685
Servorecorder, 456
Sidereal time, 158
Signal analysis:
  basic theory, 107
  histograms, 114
Signal averager, transfer function, 149
Signal averaging, 116
Signal generator:
  amplitude modulators, 584
  microwave, 576
  stability, 583
  standard, 579
Signal source, sinusoidal audio-frequency, 319
Sinad sensitivity, 519
Sine-wave synthesis, 345
Single sideband, angle-modulation, 509
Slotted line, 655
Smearing function, 143
Smith chart, 696
SMPTE method, 504
Society of Motion Picture and Television Engineers (SMPTE) method, 504
Solid-state amplifier, 550
Solid-state oscillator, 565
Solid-state sources, 572
  microwave, 556
Spectral analysis:
  by digital methods, 126
  power, 134
Spectrum analyzer, 616
  noise, 628
  real-time, 635
  superheterodyne, 620
  with tracking generator, 635
  trf, 633
Spectrum averaging, 151
Square-wave testing, 44
  effect of low-end cutoff on, 57
Stability of physical systems:
  frequency multiplier, 564
  oscillator, 578
  signal generator, 583
Standard:
  atomic, 12
  prototype, 11
  time, 157, 162
Standard deviation, frequency measurement, 164
Stationary process, 114
Step response, 48
  examples of, 50
Susceptibility:
  conducted, 515
  electromagnetic, 514
Sweep generator, microwave, 588
Switched sweep, 387
Synchronous detection, 257
Système International d'Unités, 12

Test signal:
  pseudorandom, 93
  random, 91
Test systems, automatic, 716
Thermistor, 600
Thermocouple, 596, 712
Time:
  atomic, 158
  definitions of, 157
  ephemeris, 157
  mean solar, 158
  sidereal, 158
  standard, 157, 162
  universal, 158
Time-base waveform, 385
Time-delay differential, 528
Time-domain reflectometry, 46, 61, 408
Time-signal broadcasts, 158
Time window, 126, 127
Transfer function, 481
  signal averager, 149
Transfer gain, 481
Transfer impedance, 268
Transformer-ratio-arm bridge, 296
Transient response of galvanometric recorder, 430
Transistor amplifier, 554
Transistor chopper, 207
Transmission measurement, 695
Transmitter, 507
  amplitude-modulation, 508
  commercial specifications, 539
  federal and military specifications, 540

Transmitter (*Cont.*):
keying waveshape, 534
noise-loading, 525
residual noise, 533
Trigger generator (cathode-ray oscilloscope), 389
Tunnel-diode amplifier, 551, 553
Turnaround time, 535

Units:
electrical, 7
electromagnetic (emu), 7
electrostatic (esu), 7
history of, 6
International System of, 12
definitions of, 12
Universal time, 158

Variable persistence in cathode-ray tube, 370
Variance, 84
in digitizing, 123
Vector impedance meter, 281, 310
Vector voltmeter, 29
Voltage measurement, alternating-current, 236
Voltage-to-frequency conversion, 221
Voltmeter:
averaging, 239
effects of distortion, 242
peak-reading, 245
root-mean-square, 250
sampling, 255

Warshawsky bridge, 286
Wave analyzer, 632
Wavemeter, 611
Wayne Kerr radio-frequency bridge, 305
Wheatstone bridge, 282
White noise, 71
Whitney, Eli, 5
Wien bridge oscillator, 331
Q of, 333
Writing speed (cathode-ray tube), 363

X-y plotter (oscilloscope), 411

OTHER McGRAW-HILL INTERNATIONAL STUDENT EDITIONS IN RELATED FIELDS

Alvarez: INTRODUCTION TO ELECTRON DEVICES
Auslander: INTRODUCING SYSTEM AND CONTROL
Bartee: DIGITAL COMPUTER FUNDAMENTALS, 3/e
Belove: DIGITAL AND ANALOG SYSTEM, CIRCUITS, AND DEVICES
Bowles: ANALYTICAL AND COMPUTER METHODS IN FOUNDATION ENGINEERING
Davis: INTRODUCTION TO ELECTRONIC COMPUTERS, 2/e
Davis: COMPUTER DATA PROCESSING, 2/e
Director: INTRODUCTION TO SYSTEM THEORY
Dorfman: LINEAR PROGRAMMING AND ECONOMIC ANALYSIS
Gass: LINEAR PROGRAMMING, 4/e
Ralston: INTRODUCTION TO PROGRAMMING AND COMPUTER SCIENCE

07-085544-7

electronic measurements and instrumentation

OLIVER ◆ CAGE